PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

ESCUELA DE SISTEMAS

"DESARROLLO DE UNA GUÍA METODOLÓGICA SOBRE MINERÍA DE DATOS"

AUTORES:

OSCAR JOSHUE CÓRDOVA GALLEGOS

CARLOS PATRICIO ROSALES GALLARDO

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS Y COMPUTACIÓN

DIRECTOR:

ING. ALFREDO CALDERÓN

QUITO, ENERO - 2017

Ing. Alfredo Calderón

Director de Tesis

Ing. Fabián De La Cruz

Corrector de Tesis

Ing. Damián Nicolalde

Corrector de Tesis

Dedicatorias

A Dios por darme la oportunidad de haber llegado a este punto de mi vida y haberme bendecido con salud para así lograr alcanzar mis objetivos, a su vez para proponerme nuevas metas, considerando diariamente su amor infinito, ya que él se merece todo lo que logremos en la vida.

A mi padre José Córdova por brindarme la oportunidad de poder ampliar mis conocimientos y oportunidades dándome todo su apoyo, por medio de consejos sobre la perseverancia y la constancia con lo que lo caracteriza a él, por cual es la mayor influencia y mi ejemplo a seguir como persona y profesional, pero sobre todo a su amor.

A mi madre Vilma Gallegos por ser ese pilar fundamental he importante de mi vida, el cual ha estado constantemente a mi lado apoyándome, por medio de la motivación constante que me ha permitido ser una persona honesta, pero sobre todo agradezco su amor incondicional.

A mis hermanos Scarleth Córdova y Joshua Córdova que siempre han estado pendientes en todo momento y sobre todo poder contar con ellos en los momentos difíciles para así superar las adversidades juntos, ellos son y serán lo más importante en mi vida.

Al Ing. Alfredo Calderón, Ing. Fabián De La Cruz y al Ing. Damián Nicolalde, quienes nos brindaron su apoyo constante para la realización de la presente tesis y así encaminarnos en la culminación de nuestros estudios; les agradezco de corazón por haber compartido con nosotros sus experiencias en la vida laboral y así impulsarnos o encaminarnos a nuestra formación profesional.

A mis amigos, por el apoyo, la amistad y la sinceridad que siempre nos mostramos mutuamente en nuestra formación profesional, ustedes también son parte de este logro: Juan Prado, Christian Vega, Johnny Arias, Diego Ponce, Cristhian Peñafiel, Steven Coronado; pero sobre todo a mi compañero de disertación Carlos Rosales por su entrega y dedicación para la generación de la presente Guía Metodológica.

Oscar Córdova

Dedico esta tesis a mis padres Ángel Rosales y Lupe Gallardo, a mis hermanos Vanessa y Mauricio, que, sin su apoyo incondicional durante toda mi vida, así como afecto, no habría llegado a este punto culminante de mi vida universitaria.

A mis amigos y compañeros quienes me apoyaron durante toda mi etapa universitaria, especialmente a Vanessa Soria, Juan Prado, Pablo Almeida, Johnny Arias, Steven Coronado, Juan Canelos, Pierre Quiteaquez, Christian Vega. Su apoyo fue vital durante mi vida universitaria.

Carlos Rosales

Contenido

PRÓLOGO

En el siguiente trabajo de disertación se presenta el desarrollo de una guía metodológica la cual se enfocará en la minería de datos. Como parte fundamental e inicial previa a la generación de la guía, se revisará la teoría en la que va enfocado a los conceptos básicos, a la normalización de los datos, el origen de la información y su evolución, los cuales serán de suma importancia para el entendimiento e interpretación en la aplicación de minería de datos.

El primer capítulo se describe sobre los datos y bases de datos. En donde se revisará sobre el origen de la información y el porqué de la misma. Otro de los temas a tratar será sobre los datos, la calidad de los datos, en donde se involucra varios tipos de metodologías y técnicas que son esenciales a seguir para el mejoramiento continuo. Por lo que es importante que el usuario final comprenda sobre los conceptos previos del cómo se genera valor y el cómo se llega a una reducción de costos. Adicionalmente se tratará sobre la normalización de los datos y su clasificación. Finalmente, en este capítulo se hablará sobre las diversas entidades que son capaces de almacenar grandes cantidades de información o datos.

El segundo capítulo se revisará los conceptos de la Minería de Datos. Comenzando por su historia tomando como referencia el origen, conceptos básicos y el propósito del minado de datos. Se establecerá las definiciones básicas de Bases de Datos Operativas (OLTP), en la que se estudiará la arquitectura del almacenamiento de datos y el análisis predictivo. Adicionalmente se revisará los Principios de Minería de Datos en donde se considera la reiteración y la temporalidad. Finalizando con el capítulo se estudiará las diversas Técnicas y Metodologías de la Minería de Datos que son KDD y OLAP.

En el tercer capítulo se desarrolla la Guía Metodológica para la Minería de Datos. En el cual se detallará una serie de pasos a seguir para realizar el minado de datos y el por qué de la realización del mismo.

En el cuarto capítulo se procede al realizar un ejemplo práctico donde el principal objetivo es presentar el procedimiento de la minería de datos de forma detalla, el cual será realizada siguiendo la guía, en la que se propone el uso de diversos tipos de herramientas de software y métodos especializados en este campo.

En el quinto capítulo se determinará las conclusiones y recomendaciones los cuales serán propuestos en el desarrollo de la teoría y en la generación de la guía metodológica aplicada en alguna de las herramientas de software propuesta, con la finalidad de

determinar diversas ventajas o desventajas sobre los procesos realizados para la generación de minería de datos.

INTRODUCCIÓN

La Minería de Datos se ha convertido en una parte esencial para las empresas públicas o privadas, organizaciones e instituciones educativas que administran, manejan y tratan con grandes cantidades de información por lo que es considerado como una parte fundamental para la Ciencia de los Datos.

El desarrollo constante de la tecnología permite que todas las organizaciones manejen o procesen grandes cantidades de información dando como resultado el análisis y la creación de estrategias de negocio para una mejor toma de decisiones. Por lo que es necesario el conocimiento de herramientas y entender la complejidad del manejo de datos, para la creación de posibles soluciones a diversos problemas que puedan ser presentadas en cualquier tipo de organización, generando así la mejor ruta basándose en datos históricos y proponer la mejor solución. Caso contrario, si no se posee datos históricos porque es un nuevo problema se procede a generar nuevas soluciones originando nuevas posibles rutas que ayuden para la toma de decisiones.

1. CAPITULO 1: DATOS Y BASES DE DATOS

En el siguiente capítulo se desarrollará las diversas definiciones que se desea considerar y tomar en cuenta para un mejor entendimiento. En la presente disertación se tomará como punto de partida sobre los datos, la evolución de los mismos por medio de la historia, sus principales conceptos y la complejidad al usarlos.

1.1. DATOS Y ORIGEN DE LA INFORMACIÓN

1.1.1. Datos

Es un conjunto que es reflejado de forma cualitativa o cuantitativa en la que su representación simbólica puede optar ser numérica, algorítmica, espacial, instrucciones, alfanumérica, cifras, etc., dicha agrupación puede seguir cualquier tipo de orden ya que se los considera como Datos Aislados[1], dicho conjunto de datos deben seguir un proceso de ordenamiento para que puedan ser interpretados como información, en la que describe la experiencia o las observaciones y así puede tener cierto tipo de valor instruccional[2].

Los datos representan un hecho o entidades, que pueden asociarse mediante el contexto para obtener información. La información es una constitución de un conjunto ordenado y supervisado de datos, que permite formar un mensaje y la toma de decisiones para resolver problemas guiándose en la base del conocimiento.

La base del conocimiento es el conjunto de elementos de conocimientos, que fueron recopilados como reglas o hechos, en la que se encuentra organizada y almacenada dentro de una estructura de información. En la

[1] Datos Aislados: es un concepto del conjunto de datos que no siguen ningún tipo de regla específica, los cuales necesitan o requieren un proceso de ordenamiento para que dichos datos pueden convertirse en Información, la cual ya puede ser interpretada.

[2] Instruccional: es la agrupación de diversas cantidades de palabras en la que tiene como propósito el establecimiento de reglas para que puedan ser interpretadas por cualquier persona.

que se puede acceder, recolectar, almacenar y recuperar la información que haya sido computarizada.

La información es un recurso que mediante un conjunto de datos puede representar diversos modelos del pensamiento humano, porque los datos pueden ser apreciados por los sentidos, en los que integran y forman información que es utilizada para generar conocimiento.

1.1.2. Origen de la información

En el transcurso de la historia la accesibilidad y el almacenamiento de los conjuntos de datos han ido variando según el período de tiempo y el avance tecnológico. Por lo cual nace la necesidad de almacenar y archivar grandes cantidades información de datos, para así poderlas manipular o acceder a ellas.

En la Figura 1.1. se ilustra el proceso y causa que ayudó a obtener el Origen de la información.

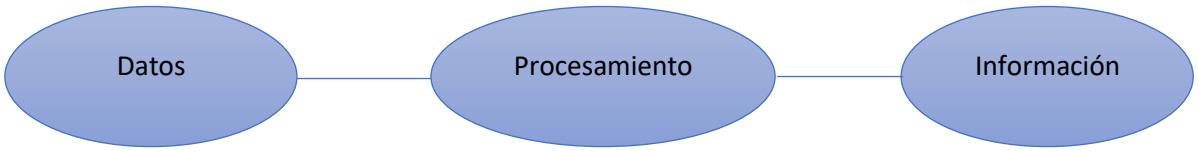


Figura 1.1. Origen de la Información, Elaborado por: Oscar Córdova y Carlos Rosales

La información que se obtiene y se almacena es de suma importancia, porque nos permite obtener una gran cantidad de información, la cual cubriría nuestros conocimientos en base a lo que necesitamos o buscamos saber y así interpretar mejor el entorno que nos rodea permitiéndonos tomar decisiones de acuerdo a los datos adquiridos.

A través del tiempo la información ha pasado por diferentes cambios que fueron necesario entre los cuales encontramos:

- Entre los siglos V y X, la única forma de obtener información era mediante las bibliotecas en la que se encontraba todo tipo de registros, cuyo acceso era limitado y en la cual toda la información era registrada manualmente. En este período se puede decir que

recopilaban la información de los habitantes de cada uno de los pueblos y sobre cómo eran las cosechas.

- Durante la edad media, con la creación de la imprenta, y la necesidad de registrar grandes cantidades de información en los libros, se distribuyeron varios de ellos así haciendo más accesible la información para las personas comunes.
- En el siglo XX, la televisión, logró que la información tenga un gran impacto en los medios de comunicación.
- En los comienzos de la era de la comunicación electrónica la palabra información fue definida en terminologías desde un punto de vista científico, dicha palabra fue definida por Jeremy Campbell[3].
- La humanidad entra en la era digital mediante la aplicación de los transmisores y el sistema binario, para representar, transmitir y personalizar la información.
- Se ven en la necesidad de almacenar grandes porciones de datos por lo que recurren a guardar, y almacenar la información, estas eran recopiladas en las que fueron las primeras computadoras.
- Nace la ARPANET[4] que ayuda a establecer las primeras conexiones entre computadoras.
- En la actualidad, se ha propuesto la globalización de acceso a enormes volúmenes de información existentes en medios más complejos y con capacidades de almacenamiento exponenciales

1.2. LA CALIDAD DE LOS DATOS Y SU COMPLEJIDAD

1.2.1. Calidad de los datos

[3] Jeremy Campbell: Fue un escrito de origen Inglés, el escribió un libro llamado la Teoría de la Información.
[4] ARPANET: Es la primera conexión que se establece entre computadoras, es lo que ahora lo llamamos Internet.

La calidad de datos es el conjunto de procesos, técnicas y metodologías, que tienen como finalidad el mejoramiento continuo sobre la eficiencia en el manejo de los datos existentes, el proceso de calidad de los datos para que sea eficaz tiene como prioridad el ser entendible y repetible para el usuario final, sea este una empresa o una organización corporativa dentro del ámbito público o privado.

Una muy buena calidad de datos es de suma importancia y decisorio ya que se vuelve un activo corporativo en donde la calidad genera consecutivamente un valor sea este en la evolución, en la gestión y en la administración de los recursos disponibles dentro de una organización para la reducción de costos y así obtener resultados beneficiosos para la misma.

Para la calidad de los datos es necesario que sigamos los siguientes puntos:

- Prevenir y evitar la duplicación de los datos [5]
- Los datos deben estar bien redactados para mantenerlos limpios y libre de errores de sintaxis[6] para el uso posterior en todas las aplicaciones que vayamos a utilizar.
- Impulsar al mejoramiento continuo[7] a partir de datos fiables como referencia de la base del conocimiento para un nuevo versionamiento.

Para la realización de un proceso de calidad, es de suma importancia tener el conocimiento de que exigencias o requerimientos son necesarios en el estado de los datos; los cuales son relacionados en un punto de partida, en los que son clasificados por su perfilamiento del impacto, estos son puntos necesarios para una corrección en casos de existir problemas de calidad.

[5] Duplicación de los datos: es un proceso que consiste en la duplicación de los datos desde un lugar a otro dentro de una red almacenamiento.

[6] Sintaxis: es la forma de cómo se combinan las palabras para la elaboración de oraciones y expresar diversos conceptos de una manera coherente.

[7] Mejoramiento Continuo: es conocido como un proceso en el que su principal enfoque es hacer más efectivo, eficiente y adaptable cualquier necesidad, en la que refleja la calidad del producto para hacerlo más competitivo a un largo tiempo.

Para estos tipos de errores de calidad son necesario unos puntos de control que servirán como indicadores para proseguir con la corrección de dichos errores, los cuales son conocidos como “Las seis dimensiones de la calidad de datos”, qué se presentan a continuación:

- Completitud[8]: ¿Existen campos en blanco? ¿Los datos pueden ser utilizables?
- Conformidad: ¿Los datos están en un formato estandarizado?
- Consistencia: ¿Consta de información contradictoria?
- Precisión: Los datos son comparados con una fuente de referencia, para confirmar que puedan ser utilizados
- Duplicación: ¿Existe información que posee un mismo formato y se encuentre en una misma tabla?
- Integridad: ¿Toda la información presente en la tabla puede ser utilizada?

En la Figura 1.2. se presenta un ejemplo sobre la Calidad de los Datos.

[8] Completitud: hacer referencia si está completo, esto quiere decir si las formulas están lógicamente validadas en un sistema.

ID	Nombre	Tipo_Persona	Fecha	Status	Direccion 1	Direccion 2
541	Carlos Rosales	Persona	23-Ene-12	Activo	10 de Agosto	Riobamba
542	Juan Prado	Persona	11-Jul-12	Activo	Las orquideas	Ambato
543	Oscar Cordova	Persona		Activo	Los girasoles	Ambato
544	Steven Coronado	Persona	23-May-12	Baja	24F	Quito
545	Puce	Empresa	6-Feb-12	Activo	Olmedo	Quito
546	Microsoft	Empresa	23-Mar-12	Activo	9 de Octubre	Guayaquil
547	TATTA	Empresa	18-Apr-12	Activo	Juan Tangamarengo	Guayaquil
548	Vanessa Soria	Persona	11-Jun-13	Activo	La independencia	Calderon
549	Movistar	Persona		Activo	La prensa	Quito
550	Supermaxi	Empresa	30-Sep-12	Activo	Brasil	Quito
551	Karina Salgado	Persona		Baja	Ventimilla	Quito
552	Baby Crazy	Empresa	4-Oct-12	Activo	Orosco	Riobamba
553	Nintendo	Empresa	3-Dec-12	Activo	Calle 17	Colombia
554	Oracle	Empresa	1-Feb-13	Activo	Gonzalo Pizarro	Quito
555	Juan Canelos	Persona	23-Nov-12	Baja	Gonzalo de Vera	Tumbaco
556	Carlos Patricio Rosales	Persona	23-Ene-12	Activo	10 de Agosto	Riobamba
557	Katerine Soria	Persona	14-Jun-12	Activo	La Occidental	Quito

Duplicidad: Se repite el registro de Carlos Rosales

Completitud: Faltan las fechas

Consistencia: Movistar no es una persona

Integridad: no se determina si son familiares

Precision: Los datos no son los correspondientes

Conformidad: La dirección no cumple con los estandares

Figura 1.2. Calidad de los datos, Elaborado por: Oscar Córdova y Carlos Rosales

1.2.2. Pasos a seguir para la Calidad de Datos

Perfilamiento:

Radica en la utilización de algoritmos para identificar los diversos tipos de contenido de los campos como:

- Perfiles de texto para nombre, direcciones y otros campos de texto (Sanchez, 2013)
- Perfiles de carácter para cédulas de identidad, teléfonos y otros (Sanchez, 2013)

E identificar los posibles problemas que podrían impedir el correcto uso de los datos.

- ¿Los campos cuentan con integridad suficiente?
- ¿Los campos contienen valores válidos y coherentes?

- ¿Qué procesos de estandarización / limpieza requieren los campos?
- ¿Cuáles reglas son efectivas?

Los resultados obtenidos son una serie de reportes que informan los problemas con los campos.

Estandarización / Normalización:

Es el proceso de eliminar o etiquetar los problemas detectados para tener una base de datos apta para la utilización en la parte práctica de la presente guía, los campos se ven expuestos a normas impuestas por el usuario para eliminar inconsistencias[9] identificadas en el perfilamiento de datos.

Se realizan las siguientes actividades:

- Eliminación de ruido
- Análisis de datos
- Estandarización de términos
- Obtener valores faltantes

1.3. NORMALIZACIÓN DE LOS DATOS

La normalización de los datos es un proceso mediante el cual una base de datos trata de evitar redundancia llegando a una estructura más funcional.

Se puede decir también que la Normalización de Datos consiste en la aplicación de reglas a las relaciones derivadas del proceso de conversión de un modelo entidad - relación a un modelo relacional.

La normalización se clasifica en objetos, elementos, relaciones y sus formas. Todo se basa en las características que cada una de las tablas posee.

[9] Inconsistencia de datos: es cuando en la base de datos existe algún tipo de incongruencia, como por ejemplo la redundancia o duplicación de la información.

Existen tres formas normales las cuales fueron creadas por Edgar F. Codd., básicamente su enfoque era el cubrir gran parte de las necesidades de la mayoría de las tablas de una base de datos. Se puede mencionar que una o todas las tablas de una base de datos se encuentran en una forma normal N.

Las formas normales generalmente ayudan a eliminar todas las inconsistencias y redundancias para que pueda existir flexibilidad en el proceso de diseño de las tablas, así permitiéndonos crear y mejorar una base de datos funcional y eficiente, evitando el perjudicar el performance por una arquitectura mal diseñada.

Generalmente en el ámbito práctico se desarrolla un modelo lógico[10] inicial por parte de los diseñadores de bases de datos para el mapear un diagrama[11] E-R de una base de datos relacional a un conjunto de relaciones de otro modelo conceptual.

1.3.1. Primera forma normal

La primera forma normal se trata de extraer todos los atributos del conjunto de entidades del modelo conceptual, es decir que cada uno de los atributos en cada una de las filas o celdas de la tabla, solo contienen un valor.

En la primera forma normal los atributos son atómicos, lo que significa que no se permiten campos repetidos, grupos, listas y conjuntos en el dominio, ya que los valores no deben descomponerse más allá de la tabla lo que nos permite evitar la duplicidad de la información.

Todos los valores de entrada en cualquier columna deben ser de la misma clase, dichos valores deben ser únicos ya que son un identificador de dicho valor, el orden de los valores no influye en la tabla.

[10] Modelo Lógico: es un tipo de modelamiento que se caracteriza porque su enfoque principal está ligado a las operaciones más que a su descripción de una realidad; detalla un esquema lógico del modelo conceptual.

[11] Diagrama de flujo de datos: es una representación de tipo gráfico para visualizar el proceso o flujo de datos y así dar soluciones a problemas de un sistema y las entidades externas.

1.3.2. Segunda forma normal

La segunda forma normal trata sobre la dependencia funcional que tienen los atributos de una tabla con la clave primaria, es decir cada atributo debe poder identificarse mediante la clave primaria y no como un atributo independiente.

Por lo tanto, en una tabla deberá existir únicamente una clave primaria con la cual se podrán establecer relaciones con otras tablas, dependiendo de la relación existente deberá ser necesario crear tablas débiles que permiten manejar dicha relación mediante las claves primarias de las tablas involucradas.

Para encontrarse dentro de la segunda forma normal es necesario que la tabla cumpla con las condiciones previas y obligatoriamente se encuentre en la primera forma normal

1.3.3. Tercera forma normal

La tercera forma normal, nos indica que dentro de una tabla no puede existir dependencia mediante la propiedad transitiva, es decir si existe la relación de dependencia A → B y B → C, entonces A → C. Cada uno de los atributos de la tabla debe ser únicamente dependientes de la clave primaria y de ningún otro atributo.

Esta forma normal es necesaria ya que si existen dependencias de manera transitiva pueden presentarse problemas al momento de realizar acciones como inserción, actualización y borrado de datos. Se pueden generar conflictos de información, así como duplicidad en las tablas.

Para declarar que una tabla se encuentra en la tercera forma normal, es necesario que cumpla las condiciones anteriores, es decir debe encontrase en primera y segunda forma normal.

Actualmente existe la regla denominada forma normal Boyce-Codd, siendo esta una forma mejorada de la tercera forma normal, ya que Boyce-Codd puede aplicarse a tablas que contienen más de una clave candidata, cosa que era imposible para la tercera forma normal.

Boyce-Codd se aplica directamente a las relaciones existentes, antes que a la tabla como tal, y dice: “Una relación está en forma Boyce-Codd sí, siempre que existe una dependencia funcional X → A, entonces X es una súper clave[12]”.

1.4. BASES DE DATOS Y SUS TIPOS

1.4.1. Bases de Datos

Son entidades capaces de almacenar datos que guardan alguna relación, de manera estructurada para poder utilizarla después. Debido a este concepto de poder acceder a la información, nos referimos a redes lo que le da el nombre de sistemas de información.

Debido a la complejidad que representaba el manejo de grandes cantidades de información, rápidamente se crearon los SGBD[13] (Sistemas Gestores de Bases de Datos). Los cuales permitían un acceso simple y una capacidad de administración de las bases de datos. Los usuarios de las bases de datos podrán ingresar mediante los SGBD para visualizar y modificar data, dependiendo de los niveles de permisos con los que cuenten, estos datos podrán ser utilizados por varios usuarios a la vez.

1.4.2. Características de una base de datos

La mayoría de bases de datos cumplen las siguientes características:

- **Independencia física:** El nivel físico puede ser modificado sin afectar el nivel conceptual.
- **Independencia lógica:** El nivel conceptual puede ser modificado sin afectar el nivel físico. Se podrán introducir mejoras sin afectar la experiencia del usuario

[12] Súper clave: está compuesto por uno o más atributos que nos ayudan a la identificación de una entidad única en un conjunto de entidades.
[13] SGBD: Sistemas Gestores de Base de Datos

- **Facilidad de uso:** Un usuario que no tenga conocimiento de la base deberá ser capaz de realizar sus consultas.
- **Acceso rápido:** La base de datos deberá ser capaz de presentar los resultados en el menor tiempo posible, para mejor esta característica se utilizan algoritmos de búsqueda que acortan el tiempo de respuesta.
- **Administración centralizada**: Los SGBD deberán permitir la actualización o manejo de datos de forma centralizada.
- **Redundancia controlada:** El SGBD deberá ser capaz de evitar la redundancia siempre que sea posible, para optimizar los recursos de la base de datos.
- **Verificación de integridad:** Los datos deberán ser coherentes
- **Uso compartido de datos:** Varios usuarios deberán ser capaces de acceder a la base de datos al mismo tiempo.
- **Seguridad de los datos:** El administrador de la base de datos deberá ser capaz de administrar los derechos de acceso de los usuarios a los datos.

En la Figura 1.3. se presenta las características de una base de datos que se deben cumplir.

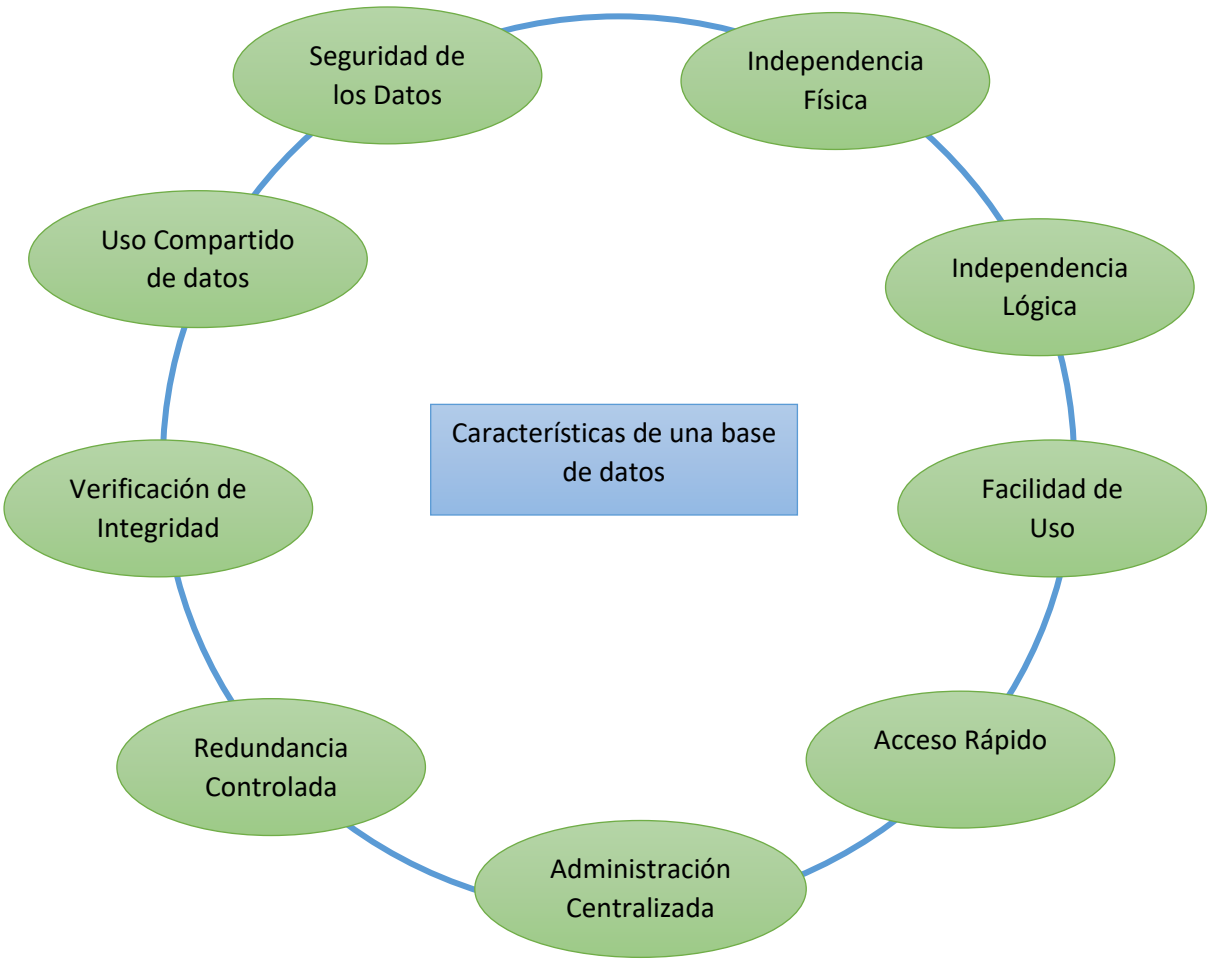


Figura 1.3. Característica de una base de datos, Elaborado por: Oscar Córdova y Carlos Rosales

2. CAPITULO 2: MINERÍA DE DATOS

En este capítulo se considerarán las diversas definiciones acerca de la minería de datos tomando en cuenta desde su origen, su evolución y el propósito del mismo, por lo que es fundamental aclarar ciertos conceptos básicos que serán de suma importancia para el aprendizaje.

La minería de datos es un mecanismo de exploración que permite el descubrimiento de información por medio de un conjunto de tecnologías y técnicas que permiten explorar información valiosa en grandes cantidades de datos. Sigue un conjunto de patrones o reglas que permiten cumplir con el contenido, el cual se encuentra guardado o respaldado en algún repositorio[14] de datos y así cubrir la necesidad del usuario final.

La minería de datos para realizar su proceso de búsqueda en la que utiliza una serie de algoritmos los cuales trabajan conjuntamente con técnicas de estadística e inteligencia artificial[15] y hace el uso de redes neuronales[16].

La minería de datos se la puede poner en práctica siempre y cuando existan base de datos de gran volumen, por lo que es necesario el conocimiento previo del proceso de minado de datos, como el entorno al que se encuentra aplicado el mismo. El objetivo que tiene como prioridad es el aprovechar al máximo el valor de la información buscada y así poder usar los patrones preestablecidos para generar un mejor entendimiento y así permitiéndonos realizar una toma de decisiones confiables.

La gran cantidad de información que es generada puede ser manipulada, administrada, controlada, optimizada, planificada y realiza predicciones, lo cual nos permitirá realizar una toma de decisiones en ámbito que este aplicado la minería de datos.

Existen algunos factores que son influyentes que nos permiten generar grandes cantidades de información:

[14] Repositorio: es un lugar en donde se respalda todo tipo de información digitalizada la cual se encuentra almacena en una base de datos.
[15] Inteligencia artificial: conocida también como AI son sistemas expertos que fueron programado para que se encarguen de imitar las acciones o el pensamiento humano.
[16] Redes neuronales: fueron realizadas en base al comportamiento de las neuronas de los seres humanos por lo que simula modelos de tipo artificial que tienen como propósito el resolver problemas de alto grado de dificultad por medio de algoritmos convencionales.

- La creación de nuevas tecnologías son factores para una mayor capacidad en los procesadores.
- Un nivel alto de confiabilidad y mayor velocidad en la transmisión de los datos.
- Administradores de bases de datos con mayor capacidad y más poderosos.
- Los sistemas de almacenamientos pueden ser de bajos costos.

2.1. HISTORIA Y EVOLUCIÓN

2.1.1. Origen

En el transcurso de la historia el hombre ha extraído información de una gran variedad de datos en varios campos usando una gran cantidad de metodologías y algoritmos matemáticos.

La extracción y recolección de dichos datos permitía generar mayor conocimiento acerca del campo que se desea emplear dichas metodologías matemáticas.

Con el avance tecnológico se abrieron nuevos horizontes para el ser humano por lo que se empezó a usar ordenadores que sirvieron de apoyo para la realización de tareas, dando así el surgimiento del aprendizaje de máquina ("Machine Learning")[17].

En el ámbito empresarial la minería de datos ha tenido una gran participación para la toma de decisiones, por lo que usan base de datos para la recopilación de información y así almacenar y gestionar grandes cantidades de datos tanto actuales como históricos. De igual forma permite la recuperación de información de grandes cantidades de datos.

Por ejemplo, el manejo adecuado o el uso de patrones en la minería de datos ayudaría a la evaluación, de diversas posibilidades, como de

[17] Machine Learning: el aprendizaje de máquina es aquello que le permite aprender a la computadora utilizando varias herramientas de software, para generar un análisis de datos en donde se observan los comportamientos que se generan de la información no estructura.

donde o cuando se puede impartir un negocio, que tipo de clientes podríamos obtener, el tipo de publicidad, etc.

2.1.2. Historia

En los años 60s los datos eran manipulados por metodologías estadísticas y matemáticas, esto era conocido como (data fishing)[18] o data archaeology, en la que realizaban la búsqueda de la información sin tener mayor conocimiento sobre las correlaciones en base de datos.

La minería de datos nace a partir de los grandes volúmenes de información que obtenían los investigadores, por lo que en la década de los 80s un grupo de investigadores entre los que se encontraba Gio Wiederhold[19], Robert Blum, Gregory Piatetsky[20] y Rakesh Agrawal[21] asentaron nuevas terminologías para el manejo de grandes cantidades de información llamado así "Data mining". En esta década solo existían pocas empresas que hacían el uso de tecnología vinculada al manejo de datos.

Con el desarrollo de las nuevas tecnologías los proveedores de los Sistemas de Manejo de las Bases de Datos, como por ejemplo IBM y Oracle desarrollaron nuevas características a sus productos de software que permitan el almacenamiento a partir de sus bases de datos estándar.

En el año 2002 ya se habían creado cerca de 100 empresas que se dedicaban a dar soluciones en más de 80 países, el cual el manejo

[18] Data Fishing: "Es el uso de la minería de datos para descubrir patrones en los datos de que puede ser presentada como estadísticamente significativo, sin elaborar primero una hipótesis en cuanto a la causalidad subyacente" (Wikipedia, 2016)

[19] Gio Wiederhold: Profesor de Ciencias de la Computación, su origen es italiano y estudio en la Universidad de Stanford, la mayor parte de su vida se dedicó a la investigación de los grandes sistemas de gestión de bases de datos.

[20] Gregory Piatetsky: Es un Científico de origen ruso, fue el co-fundador de KDD y está asociado al descubrimiento de la minería de datos.

[21] Rakesh Agrawal: Doctor en Ciencias y Computación, graduado en la Universidad de Wisconsin-Madison, ayudó al desarrollo de tecnologías, conceptos y herramientas para la minería de datos.

de datos estuvo vinculado específicamente al área académica y al empresarial.

2.1.3. Propósito del minado de datos

En la actualidad el propósito de realizar minerías de datos en parte es para generar una gran ventaja competitiva tanto para las empresas como para las instituciones educativas, que esta genera una gran cantidad de información que nutre el conocimiento y así generar beneficios a diversas escalas.

Por lo que el uso principal de la minería de datos por parte de estas instituciones es para obtener los siguientes beneficios:

- Ayuda a la optimización y el manejo de recursos de cualquier tipo de organización, la minería de datos ayuda a presentar una gran cantidad de metodologías que generan una mayor productividad y así proponer una ventaja competitiva sobre otras organizaciones.
- Realizar una clasificación de datos a partir experiencias y categorizarlas basándose en datos históricos, aquí se emplea la minería de datos basándose en datos históricos para así determinar una nueva conclusión o posible solución para algún problema. Por ejemplo, si esto es empleado en la medicina en su totalidad podemos determinar muchos de los posibles factores de un problema de salud y así generar un diagnóstico adecuado para un paciente, el cual está determinado en registros anteriores de pacientes con el mismo problema.
- Obtener predicciones de posibles comportamientos futuros, esto se basa en tomar datos e información actual o histórica que se va presentando constantemente y así predecir un supuesto comportamiento, el cual puede ayudar a una toma de decisiones para los próximos eventos.
- Reconocer la existencia de algún patrón o alguna actividad de cualquier evento pasado, en este caso la minería de datos nos

permitiría explorar algún registro histórico que cumplan con el mismo patrón y así identificarlo para generar una mejor solución al mismo.

2.2. CONCEPTOS BÁSICOS

2.2.1. Base de Datos Operativas (OLTP)

Son aquellas que tiene la capacidad de procesar grandes cantidades de información y realizar un número limitado de transacciones en línea (OLTP)[22] de forma repetitiva, este tipo de transacciones repetitivas pueden generar o requerir cambios en las tuplas de las bases de datos relacionales.

"Una base de datos como ésta se desarrolla para servir a las necesidades de información de los usuarios finales, y está diseñada para soportar sus operaciones empresariales diarias" (Ricardo, 2004)

Por lo que es necesario conocer el comportamiento de las bases de datos operativas, ya que nos permite generar un apoyo para las transacciones de los grandes volúmenes de datos que son administradas y así realizar un entregable con las posibles respuestas para los usuarios. Las bases de datos operativas siempre están en una actualización constante ya que es en tiempo real, debido a que depende de la cantidad de transacciones de la entidad o institución; de igual forma es indispensable que las operaciones como el borrado o las inserciones deben ser ejecutados de una forma espontánea y así mantener la base de datos actualizada, para generar informes sobre la situación en la que se encuentra dicha entidad.

2.2.2. Arquitectura del almacenamiento de datos

La arquitectura de datos hace referencia al cómo deben estar estructurados y diseñados los diversos métodos que sirven para el

[22] OLTP: Online Transaction Processing (Procesamiento de Transacciones en Línea).

almacenamiento de información dentro de una base de datos, el cual es de suma importancia en la mayor parte de instituciones ya que ayuda a guardar grandes cantidades de información debido a su buena arquitectura de almacenamiento, porque emplea un diseño de alto nivel para su implementación.

"Los datos se toman de las fuentes de datos, que pueden incluir bases de datos operativas múltiples, otras entradas como archivos independientes y datos ambientales como información geográfica o datos financieros." (Ricardo, 2004)

De igual forma los datos almacenados pueden ser validados e integrados para verificar el nivel de calidad del almacenamiento de información, este proceso se realiza antes de cargar la información, este proceso es considerado como la limpieza de datos.

"El proceso de carga es una transacción larga, pues por lo general está involucrado un gran volumen de datos, de modo que el sistema debe usar herramientas de gestión de transacción para garantizar recuperación adecuada en el evento de falla durante la transacción de carga" (Ricardo, 2004)

Posteriormente el almacenamiento de datos es usado para soportar las consultas OLAP[23], con la finalidad de generar posibles soluciones estratégicas que ayudan a realizar una buena toma de decisiones y así generar la información que va a ser empleada en el minado de datos. Hace uso de los diversos patrones para encontrar la información que será empleada para generar una posible solución.

A continuación, en la Figura 2.1. tenemos un ejemplo de cómo está estructurada la Arquitectura del almacenamiento de datos.

[23] OLAP: On-Line Analytical Processing (Procesamiento Analítico en Línea).

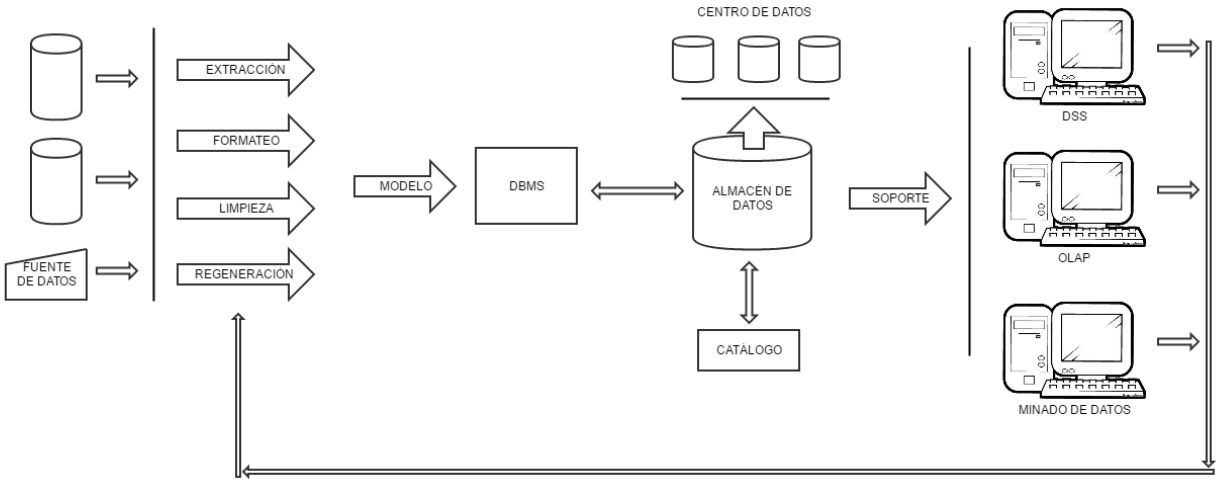


Figura 2.1. Arquitectura del almacenamiento de datos, Elaborado por: Oscar Córdova y Carlos Rosales

2.2.3. Minería de Datos (Concepto general)

"La minería de datos es el proceso por el que se extraen datos previamente desconocidos de grandes bases de datos y se utiliza para tomar decisiones en la organización" (Kantardzic, 2002).

Se realiza la preparación de datos, en la que se aplica los diversos procesos de minería de datos con la finalidad de encontrar un patrón del se obtendrá como resultado la información necesaria que ayudará a la toma de decisiones.

Existen algunos factores que son influyentes como características de la minería de datos:

- Encuentra diversos patrones que pueden ser impredecibles los cuales generan soluciones al problema.
- Un nivel alto de confiabilidad y una mayor velocidad en la transmisión de los datos es de suma importancia ya que trabaja sobre grandes volúmenes de información.
- Ayuda a la toma de decisiones dentro de una organización privada o pública.
- Los sistemas de almacenamientos pueden ser de bajos costos. Por lo que es aplicada a diversas áreas como geológicas, meteorológicas, industriales y al comercio.

2.2.4. Análisis Predictivo

"El análisis predictivo agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos" (Wikipedia, 2016)

Por lo general el análisis predictivo es usado constantemente en diversas organizaciones para la optimización de procesos, ya que analiza los diversos comportamientos y busca posibles soluciones anticipándose a los problemas generados.

Simula el comportamiento humano, por lo que utiliza soluciones basadas en el aprendizaje supervisado, en el cual se tiene dos fases fundamentales.

La primera fase es el aprendizaje en donde se dan los diversos parámetros y así se construyen los modelos de entrenamiento de las muestras de información o de datos.

La segunda fase es de comprobación, busca y analiza la información guardada en la base de datos y provee nuevas soluciones evitando la redundancia de información.

"El modelo predictivo se asocia con las técnicas de clasificación y predicción de valores. La clasificación supone determinar una clase para cada fila de la base de datos. Esto puede hacerse utilizando árboles de decisión o redes neuronales." (Davies, 1996, 2000, 2004, 2014)

Con un conjunto de información o datos de entrada se realizará el siguiente análisis, en el que este será representado en forma de un árbol de decisión[24], en donde los nodos deben cumplir una serie de condiciones las cuales se emplean para hallar una solución. Dicha solución se las representará como las ramas del árbol.

[24] Árbol de Decisión: es la representación de tipo gráfico en donde se basan las decisiones de tipo secuencial en resultados de probabilísticos. Son usados por lo general en sistemas expertos, en árboles de juego y en búsquedas binarias.

Finalmente, los nodos hojas simbolizará el rango de clases los cuales compone la fila.

Estos nodos se pueden dividir en:

- **Capa de entrada:** depende de las particularidades de los datos de entrada a la red.
- **Capa de procesamiento:** hace uso de los pesos que fueron asignados en las relaciones entre los nodos, posteriormente proceden a clasificarlos.
- **Capa de salida:** una vez clasificados en función de los pesos se visualiza los tipos de clases en el problema.

En la Figura 2.2. tenemos un ejemplo de cómo se realiza un Análisis Predictivo.

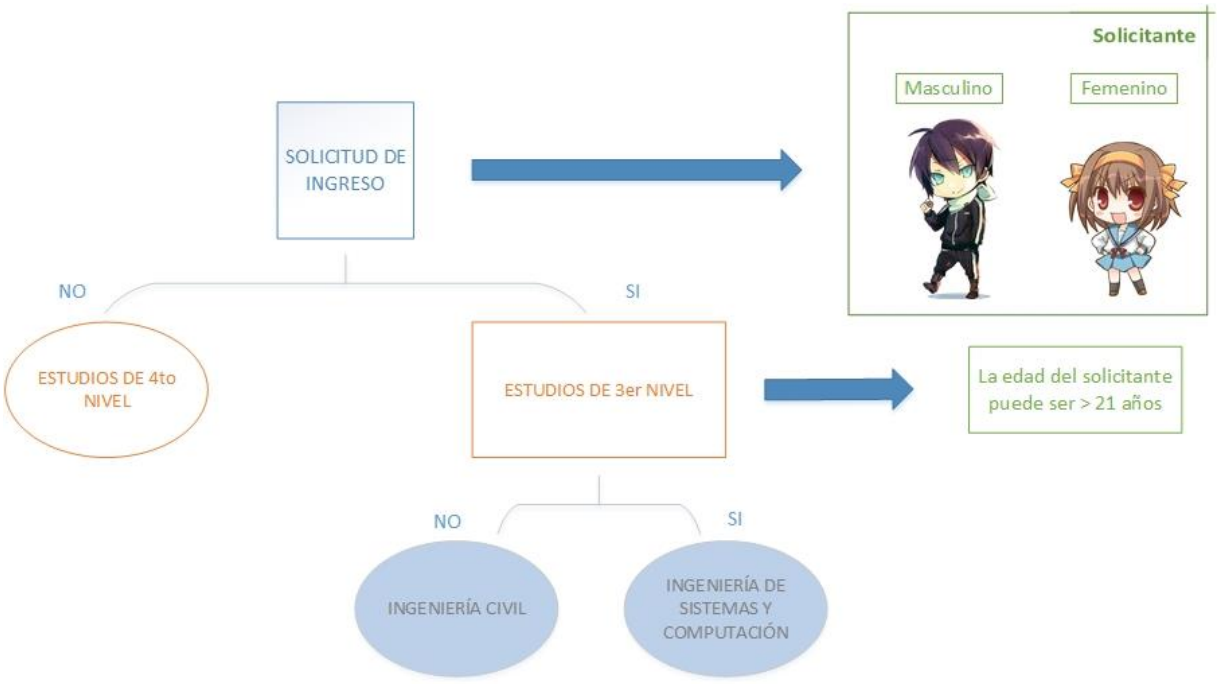


Figura 2.2. Análisis Predictivo, Elaborado por: Oscar Córdova y Carlos Rosales

2.3. PRINCIPIOS DE MINERÍA DE DATOS

2.3.1. Reiteración

Es la aplicación de un conjunto de herramientas computacionales y técnicas que se encargan de realizar diversas combinaciones en donde se toma en cuenta todos los factores para obtener un buen resultado, el cual nos ayudará a tener una mejor visión acerca de lo que ocurre y lo que esconde toda la información que ha sido almacenada en la base de datos.

2.3.2. Temporalidad

En este caso es indispensable considerar de antemano la cantidad de tiempo que puede tomar la búsqueda de un resultado, ya que se debe especular o reconocer el tipo de herramienta que se vaya a emplear. Es necesario que los datos no sean inconsistentes o demasiados pobres ya que podría no generar ninguna solución y a su vez no origine resultados que sean de interés para la organización.

Al emplear la metodología adecuada y realizando una configuración óptima de las herramientas, los patrones van a empezar a surgir y los resultados esperados se van ir obteniendo consecutivamente.

2.4. TÉCNICAS Y METODOLOGÍAS DE LA MINERÍA DE DATOS

Al referirnos de minería de datos hablamos sobre diversas técnicas estadísticas en las que se facilitan las consultas, la creación de informes, etc., por lo que es esencial considerar para la minería de datos sus técnicas importantes para su operación que son: KDD y OLAP.

2.4.1. KDD

KDD[25]: Knowledge Discovery in Databases

[25] KDD: Knowledge Discovery in Databases.

Se define al proceso como “El proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y en última instancia comprensibles en los datos” (Usama Fayyad, 1996)

Es el proceso de obtener conocimiento e información relevante y útil de la base de datos. Se utiliza cuando el crecimiento de una base de datos supera las habilidades humanas para analizar la información. Esto debido al gran tamaño de las bases de datos, la presencia de ruido, así como los datos inconsistentes y redundantes, se hacen necesarios las técnicas de pre procesamiento en las cuales se aplican minería de datos. Siendo un proceso iterativo, nos permite reconocer o determinar las diferentes relaciones existentes entre la información obtenida de la base de datos. Permitiendo la toma de decisiones en base a dichas relaciones.

El proceso de KDD está ilustra en la Figura 2.3.

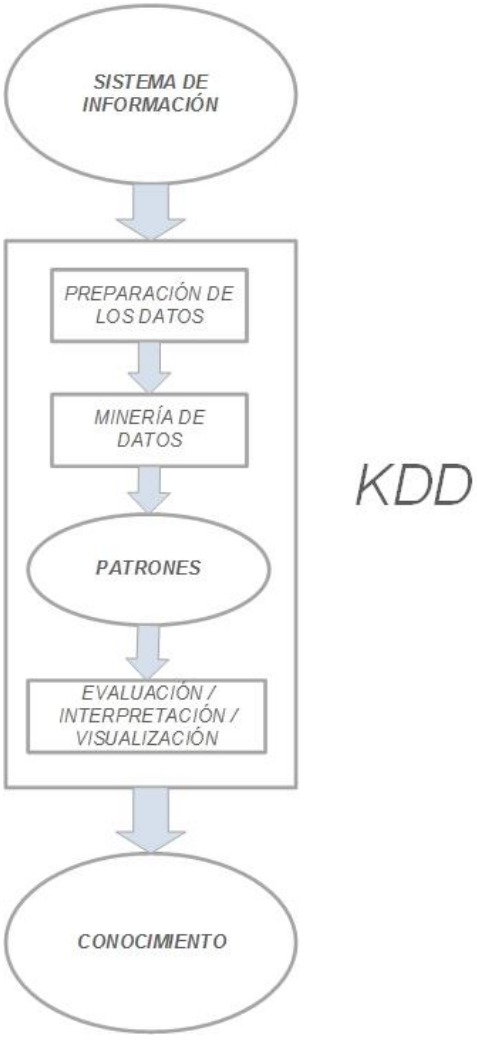


Figura 2.3. KDD, Elaborado por: Oscar Córdova y Carlos Rosales

2.4.1.1. Las etapas o fases que conforman el proceso KDD:

En este punto se hablará sobre las fases que conforman KDD que son: selección de datos, preparación de datos, data mining, interpretación y evaluación.

Selección de datos:

Esta es la etapa en la cual se determinan las fuentes de datos, así como el tipo de información que se utilizará. Dicha información se encuentra en bases de datos u otras fuentes de información ya sean externas o internas.

La fase de minería de datos dependerá mucho del tipo de fuente, así como de los usuarios que utilizan la información.

- **Picapedreros:** Realizan informes, controlan valores y supervisan la evolución de ciertos parámetros.
- **Exploradores**: Buscan patrones significativos utilizando la minería de datos.

Cuando la información es obtenida de una fuente externa es necesario determinar su tipo y utilidad, para facilitar el proceso de análisis.

Esta información puede ser demográfica, uso de internet, información de la industria o negocio, evolución económica, climatológicas, bases externas compradas a otras compañías, etc.

Preparación de datos:

El objetivo de esta fase es el obtener un conjunto de datos de calidad para utilizarlos como entradas y así obtener modelos, patrones o reglas de mayor calidad. Es decir, consiste en realizar la limpieza de datos que pueden estar incompletos (valores o atributos perdidos), presentar ruido (valores incorrectos o inesperados) o inconsistentes (valores o atributos que son diferentes a lo esperado). Dependiendo de los datos sucios estos

pueden ser simplemente eliminados, ya que mantenerlos en la base de datos puede representar un análisis inexacto o resultados incorrectos.

Las acciones que se pueden tomar en cuenta para realizar la limpieza frente a datos anómalos son las siguientes:

- **Ignorar:** Algunos algoritmos de minería de datos son robustos a los datos anómalos, es decir que el algoritmo puede manejar los posibles errores presentes en los datos.
- **Filtrar la columna:** Eliminar o reemplazar la columna, es una solución extrema pero válida, es preferible en la mayoría de los casos reemplazar la columna a eliminarla ya que pueden existir datos dependientes de dicha columna.
- **Filtrar la fila:** Sesgar los datos ya que en muchos casos datos erróneos se deben a casos especiales.
- **Reemplazar el valor:** Utilizar el valor “null” para reemplazar un valor erróneo, siempre y cuando el algoritmo de minería de datos pueda liderar con estos, en casos numéricos también se pueden utilizar máximos o mínimos dependiendo de la variable, teniendo en consideración como afectará esto al resultado.
- **Discretizar:** Transformar variables continuas a variables discretas.

En caso de datos faltantes se tomará las siguientes medidas:

- **Ignorar:** Algunos algoritmos son robustos al momento de manejar datos faltantes, como podrían ser los árboles
- **Filtrar la columna:** Eliminar o reemplazar la columna, igualmente se trata de una solución extrema, es preferible reemplazar la columna con una variable booleana indicando la existencia del valor.
- **Filtrar la fila**: Muchos de los datos faltantes se relacionan con casos especiales
- **Reemplazar valores:** A veces se puede predecir los datos a partir de otros datos, por lo que se reemplaza con valores medios.

- **Segmentar:** Las tuplas se segmentan dependiendo de los valores disponibles, se obtienen modelos para los segmentos y se vuelven a combinar.
- **Modificar la política de calidad de datos:** Esperar hasta que los datos faltantes estén disponibles.

Es necesario analizar las razones por las que se encuentran los datos faltantes, muchas veces representan características relevantes, otros valores pueden como no existir en la realidad o muchas de las tuplas faltantes pueden representar que provienen de fuentes externas.

Frente a las inconsistencias, dos valores con el mismo atributo, pero con diferente valor, es necesario unificar siempre que se puedan los registros en una única clase.

Al realizar estos procesos de limpieza sobre la base de datos, se reduce su número de datos, buscando utilizar los más relevantes con respecto al objetivo que se busca al realizar la minería de datos, así como tipos de datos específicos o un límite máximo de registros con lo cual se vuelve más efectivo el proceso de minería de datos al que se ve expuesto este grupo de información, reduciendo así considerablemente el tiempo de procesamiento en la fase de minería de datos.

Las transformaciones de datos, como la “Numerización” o “Discretización”, consisten principalmente en modificaciones sintácticas que no alteran o cambian el valor de la información, esto se realiza con el fin de crear nueva estructura de datos apropiada para el proceso de minería de datos conocida como la “Vista Minable”

Algunas de las transformaciones que se realizan son:

- **Numerización:** Se refiere a la transformación de datos nominales a datos de valor numérico, esto se realiza al utilizar una técnica de minería de datos que no acepta datos nominales.
- **Discretización:** Transformar valores numéricos a valores nominales, con el fin de establecer relaciones.

- **Aumento de dimensionalidad:** La creación de nuevos atributos sustituyendo los actuales, o añadiendo nuevos. Tiene el fin de permitir la resolución de problemas que de lo contrario serian irresolubles.
- **Reducción de dimensionalidad:** Es la creación de atributos que sustituirán los actuales, siendo estos de menor cantidad. Los patrones en muchas ocasiones no pueden formarse o detectarse debido al exceso de atributos o instancias.
- **Normalización de rango:** Se refiere a la normalización de valores numéricos, para que los atributos se encuentren dentro de la misma medida.

Data Mining

En esta fase se utiliza la vista minable, obtenida previamente mediante la preparación de datos, en la cual se aplicarán los diferentes procesos de minería de datos con el fin de obtener los patrones de los cuales se obtendrá la información útil para la toma de decisiones.

Dependiendo del tipo de búsqueda que se desee realizar existen dos diferentes tipos de minería, que son la Direct data mining (Supervisada o Descriptiva) o Undirected data mining (Predictiva).

En la búsqueda supervisada es necesario tener claro el tipo de información que deseamos obtener al final del proceso, ya que esto marcará el tipo de algoritmo que deberemos utilizar, en otras palabras, este proceso es aquel en el cual utilizaremos variables conocidas para predecir el valor de otra variable también conocida.

Mientras que en la búsqueda predictiva se utilizan variables conocidas para determinar el valor de una variable desconocida.

- Estadísticas y algebraicas: Se trata de expresar los modelos o patrones obtenidos mediante una resolución matemática ya sea como una fórmula, una función, una distribución o valores estadísticos. Usualmente este tipo de técnica se utiliza en modelos predeterminados de las cuales se pueden obtener variables o parámetros.

- Técnica bayesiana: Como su nombre lo indica se utiliza el teorema de Bayes para determinar la pertenencia a cierto grupo o clase mediante probabilidades.

  "Naive Bayes es una técnica de clasificación y predicción que construye modelos que predicen la probabilidad de posibles resultados. Naive Bayes utiliza datos históricos para encontrar asociaciones y relaciones y hacer predicciones." (Gabits, 2009)

  $$P(h|O) = \frac{P(O|h) \text{ x } P(h)}{P(O)}$$

  En la formula interpretamos que (h) es la hipótesis, (O) son las observaciones y las probabilidades condicionales corresponde a $P(h|O)$ y $P(O|h)$;en donde $P(h|O)$ corresponde a que la observación sea resultado de un experimento, el cual es (h) en la formula.

- Arboles de decisión y sistemas de aprendizaje: Son un conjunto de reglas y condiciones jerárquicas, que permiten la toma de decisiones, con tan solo verificar las condiciones que se cumplen, por lo cual los resultados son excluyen y dan como resultado una única acción a tomar. Los sistemas de aprendizaje utilizan la clasificación para determinar a qué clase pertenece el objeto, para lo cual utilizan los arboles de decisión, siendo cada uno de los nodos disyuntivos (cumple o no cumple) mediante un conjunto de condiciones excluyentes, permitiendo llegar a una única solución. Este algoritmo es conocido como algoritmo de partición o "Divide y vencerás". Hay que tener en cuenta que una vez seleccionada la partición no se podrá cambiar por lo cual es necesario un buen criterio de partición, ya que una mala elección al comienzo dará como resultado un mal árbol.

- Conteo de frecuencias y tablas de contingencia: Se refiere a contar la frecuencia con la que dos o más sucesos se presentan. Existen algoritmos que permiten este conteo cuando la cantidad de sucesos son muy grandes. Ejemplo algoritmo A priori (algoritmo usado para encontrar reglas de asociación, basado en conocimiento previo o "apriori")

- Redes neuronales:

  Se trata de una técnica que mediante el entrenamiento de pesos en nodos conectados entre sí o neuronas puede aprenderse un modelo, esto depende de la topología de la red y los pesos asociados.

- Basados en casos, densidad o distancia: Algoritmos basados en la distancia entre elementos para determinar funciones de densidad(poblacional).

Interpretación y evaluación

A partir de los resultados obtenidos de la fase anterior (patrones) se proponen hipótesis o modelos que necesitan ser evaluados y validados, que siguen el siguiente proceso:

- Primero, se valida la precisión del modelo planteado, esto se realiza comparándolo con otros modelos independientes del proyecto dando así una directiva de la dirección que se debe seguir.
- Segundo, realizar un experimento del modelo poniéndolo en práctica, con un subconjunto de usuarios o clientes que representen la población a la que se verá expuesta el modelo en un futuro, de esta manera confirmar su validez y si cumple o no con el objetivo con el que se lo desarrolló.

DATA-WAREHOUSING (Almacenes de datos)

La data warehousing es el proceso de construir una data warehouse, que es la unión de diferentes fuentes de datos heterogéneas que soporten reportes, ad hoc "*queries"* estructurales o no, y toma de decisiones. Esto involucra la limpieza, integración y consolidación de los datos.

Existen tecnologías que permiten el utilizar la información almacenada en la data warehouse haciéndolas más rápidas y eficientes. La información obtenida puede ser utilizada en los siguientes dominios, por ejemplo:

- Estrategias de Producción: Es utilizada para reposicionar los productos y el manejo de los portafolios, optimizando la cantidad de producto necesario y disminuyendo la pérdida de productos y capital
- Análisis de clientes: Se analizan las preferencias de compra de los clientes, para crear mejores estrategias de venta y publicidad, aumentan la posibilidad de ventas y los ingresos
- Análisis de Operaciones: Ayuda a mejorar el trato de los clientes, consiguiendo así mejorar la aceptación de los productos o servicios y permite la corrección de errores en las operaciones con el fin de volver más óptimo la producción.

2.4.2. OLAP

OLAP[26]: On-Line Analytic Processing

OLAP tiene como objetivo agilizar las consultas realizadas en las bases de datos, es decir crear información resumida que represente todo el contenido de la base de datos y utilizar esta información para acelerar el proceso. Aunque existe la posibilidad de realizar esto mediante consultas SQL, en muchos de los casos son muy costosas y complicadas de realizar, por lo pue representa un gasto de recursos y tiempo elevado, mientras que utilizando las herramientas de OLAP el proceso es casi instantáneo.

Los procesos estadísticos necesitan utilizar grandes cantidades de información para obtener resultados, lo que requiere manejar las agrupaciones de ciertos atributos para obtener resultados. Existen atributos de medida que son aquellos que representan cantidades o valores numéricos, como su nombre lo indica se utiliza para medir algún atributo. Por otro lado, también existen los atributos dimensionales, son las dimensiones en las que se representan los atributos de medida. Aquellos datos que pueden representar tanto una dimensión como valores de medida son los llamados datos

[26] OLAP: On-Line Analytic Processing

multidimensionales. Para entenderlo mejor utilizaremos un ejemplo de tabla de una tienda de ropa, en esta existen varios atributos como el artículo, el color, la talla y la cantidad. En dicha tabla el atributo de medida será la cantidad, ya que mide la cantidad de unidades disponibles de cada prenda, el resto de atributos son dimensionales ya que representan las diferentes dimensiones que son medidas mediante el atributo de cantidad.

Para la representación de los datos multidimensionales es necesario que expresemos la tabla de forma cruzada permitiéndonos representar los datos multidimensionales de una manera más fácil de reconocer.

Para realizar la representación de datos multidimensionales usaremos la siguiente Tabla 2.1.

	Oscuro	Pastel	Blanco	Total
Falda	8	35	10	53
Vestido	20	10	5	35
Camisa	14	7	28	49
Pantalón	20	2	5	27
Total	62	54	48	164

Tabla 2.1. Tabla de datos tridimensional, Elaborado por: Abraha, Silberschatz, Henry F. Korth, S.Sudasrshan (Fundamentos de bases de datos – Quinta Edición)

Este es un buen ejemplo donde podemos apreciar las columnas resumen que nos muestra la información necesaria de cada objeto, muchas de las tablas cruzadas tienen estas columnas. Además de permitirnos expresar los datos multidimensionales con mayor facilidad, las tablas cruzadas nos permiten añadir una mayor cantidad de atributos lo que permitirá una mejor representación de la información.

La generalización de las tablas cruzadas, que pueden ser bidimensionales a n dimensionales, es denominada como cubo de datos, donde cada una de las caras representa una dimensión de las tablas cruzadas, en nuestro ejemplo la información se representa mediante números, siendo esta, la información resultante del cruce de las dimensiones disponibles, se presenta la información en una

cara de las celdas mientras que las otras se las deja en blanco si es que son visibles, cabe destacar que aun si ninguna de las carilla es visible a primera vista igualmente contiene la información.

Una de estas celdas puede representar la agrupación de todas las celdas de dicha dimensión, lo que permite realizar agrupaciones a gran escala, así como agrupaciones de agrupaciones.

En la Figura 2.4. apreciaremos la Tabla 2.1. en forma de datos multidimensionales, en donde se ve una tabla cruzada con una gran cantidad de atributos. En donde la representación gráfica de las tablas cruzadas que pueden ser bidimensionales es denominada como Cubo de datos.

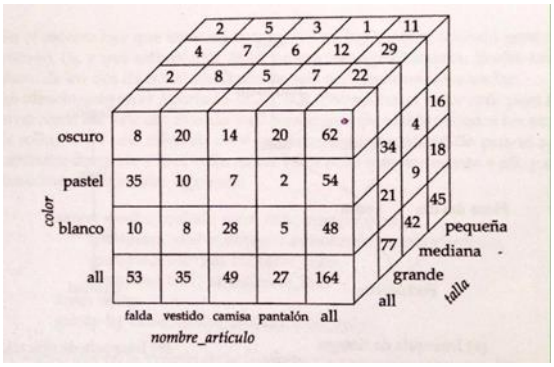


Figura 2.4. Cubo de datos tridimensional, Elaborado por: Abraha, Silberschatz, Henry F. Korth, S.Sudasrshan (Fundamentos de bases de datos – Quinta Edición)

El sistema OLAP es interactivo por lo cual permite el presentar diferentes resultados de los cruces dimensionales, aun si se utilizan los mismos atributos en varios cruces, por ejemplo, se puede presentar la información del cruce entre nombre_articulo y las tallas y a su vez el cruce de talla con color.

El procesamiento en línea permite obtener los resultados y resúmenes de la información en pocos segundos, lo que deja optimizar el rendimiento de los analistas.

Los cubos de datos también pueden ser útiles cuando se necesita tabulaciones cruzadas de varias dimensiones al mismo tiempo, un ejemplo podría ser el conocer los nombres de los artículos de talla grande con el color pastel. A estos se les conoce como cortes de

cubos ya que se asemejaría a realizar un corte en el cubo para poder ver la información correspondiente que se encuentra dentro del cubo.

Debido a que los costos de almacenamiento y mantenimiento se han reducido se crea el almacenamiento de todos los datos en un sistema unificado, se separa la información de las fuentes transaccionales. Esta tecnología facilita mucho OLAP.

Existen 4 tipos de servidores OLAP

- **Relational OLAP (ROLAP)**

  Son los servidores que se encuentran entre el back-end del servidor y el front-end del cliente. Utiliza base de datos relacionales o que soporten extensiones de bases relacionales.

- **Multidimensional OLAP (MOLAP)**

  Utiliza arrays basados en motores multidimensionales que soportan diferentes vistas de la información. Muchos servidores MOLAP utilizan dos niveles de representación del almacenaje de datos.

- **Hybrid OLAP (HOLAP)**

  Es una combinación de servidores ROLAP y MOLAP que ofrece mayor escalabilidad y rapidez de procesamiento.

- **Specialized SQL Serverx**

  Servidores avanzados con soporte de queries

2.5. HERRAMIENTAS DE MINERÍA DE DATOS

El principal objetivo de las herramientas que se usarán, es proveer un mejor enfoque y entendimiento sobre lo que representa la minería de datos, el cual será representado por alguna de las metodologías o algoritmos propuestos anteriormente para este tema.

2.5.1. PENTAHO

Es un conjunto de programas y herramientas que fue diseñada propiamente para la realización de Business Intelligence[27] la cual se centra en las metodologías y procesos para la búsqueda de mejores soluciones. Pentaho se ha empleado en muchos campos en el que se manejan grandes cantidades de información como la minería de datos, la generación de informes y en movilizar grandes cantidades de datos en ETL[28]. Pentaho ayuda a la toma de decisiones en cualquier tipo de organización, buscando así una mejor solución para dicha entidad.

Entre las principales características de Pentaho es que se centra a procesos y se orienta a la búsqueda de soluciones, por lo que es llamada como una herramienta o una plataforma enfocada a la Inteligencia de Negocios (Business Intelligence). Esta herramienta esta apta para la ejecución de las reglas del negocio, la cuales están representadas en un conjunto de actividades o procesos para así presentar una mejor solución en el tiempo que sea requerido.

Pentaho puede integrarse con cualquier motor de bases de datos sea este libre o licenciado. Los motores de bases de datos de tipo Open Source pueden ser MySQL, PostgreSQL, monetdb, etc.; mientras que en los motores de bases de datos con licencia puede ser ORACLE, IBMDB2, SQLServer, etc.

Los principales productos que ofrece Pentaho son:

Pentaho para Apache Hadoop

Pentaho Dashboard

[27] Business Intelligence: “Business Intelligence es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.” (Sinnexus, 2016)

[28] ETL: Extract, Transform and Load (extraer, transformar y cargar), es usado para el movimiento de grandes cantidades de datos o información a múltiples fuentes, en donde las bases de datos son limpiadas, reformateadas y la información es cargada con el propósito de apoyar el análisis del proceso del negocio.

Pentaho Data Mining Pentaho Reporting

Pentaho Analysis Services

2.5.2. WEKA

WEKA (Waikato Environment for Knowledge Analysis)

Es una herramienta que sirve particularmente para la minería de datos que sirve para la extracción de la información de cualquier tipo de base de datos. Entre las particularidades de esta herramienta es que es un conjunto de librerías basadas en JAVA.

La herramienta Weka tiene un gran número de algoritmos y metodologías los cuales son aplicados a grandes cantidades de datos permitiendo hacer un minado de datos por medio del pre-procesamiento de datos, los cuales son clasificados por medio de reglas y asociaciones.

3. CAPÍTULO 3: DESARROLLO DE GUÍA METODOLÓGICA

En el siguiente capítulo se explicará sobre lo que es necesario para la ejecución de minado de datos, por lo cual se presentará una guía metodológica con sus respectivos pasos a seguir.

En el desarrollo de la guía se podrá visualizar el principal propósito de la Minería de Datos la cual tiene como objetivo la extracción no trivial de información contenida en grandes bases de datos, las cuales pueden tener un sin número de información, datos o contenido como imágenes, archivos de video o música, etc. Los cuales son procesados, analizados y tratados por medio de metodologías o técnicas que ayudan a extraer la información que se desea conseguir de grandes cantidades datos.

En la actualidad la minería puede estar enfocada para varios usuarios o propietarios, sin dar importancia a la información o a qué tipo de organización pertenece, porque estas herramientas están desarrolladas con la finalidad de satisfacer las necesidades y llegar a la solución más óptima para los usuario u organizaciones sea estas pública o privada ya que no posee ningún tipo de restricción.

La guía a elaborarse está orientada principalmente a tener una mejor evidencia sobre el uso de las herramientas de minería de datos a nivel general, es de suma importancia la comprensión resultados, tomando en cuenta los principios básicos y las metodologías que se emplean en el manejo de grandes volúmenes de información, para la optimización de un mejor resultado.

3.1. GUÍA METODOLÓGICA PARA EL MINADO DE DATOS

El principal objetivo de la guía a realizarse es que la misma provea información y de una mejor orientación a las diversas entidades e individuos que van a estudiar sobre el manejo y minado de datos, el cual por medio de diversas técnicas adecuada para cada uno de las situaciones deseadas presente los mejores resultados para que la toma de decisiones sea la óptima. Teniendo en consideración que se eviten los diversos problemas que se presenta al momento de generar o realizar la

práctica, entre una de esas dificultades que se presentan está el duplicamiento de información o de datos.

Entre otros de los posibles problemas que se presentan con frecuencia dentro de las bases de datos es la falta de información o datos incompletos, los cuales cuyos valores o atributos obligatorios son faltantes. Este problema se puede dar por el mal procedimiento del manejo de los datos, la captura de los datos de forma manual, errores en la migración de datos entre diversas aplicaciones y errores que pueden presentar los equipos.

Con la minería de datos analizaremos cualquier tipo de base de datos que sea de nuestro interés los cuales por medio de patrones y metodologías aportarán un nuevo conocimiento sobre los comportamientos o las dinámicas que van a poseer los datos a analizar.

3.1.1. Pasos para realizar una minería de datos

Para realizar una minería de datos tenemos que guiarnos mediante las etapas del proceso KDD (que se menciona previamente). Para poder seguir esto, debemos tener en cuenta que información necesitamos y con la que vamos a trabajar.

3.1.1.1. Paso 1. Identificar las necesidades

El objetivo principal de la minería de datos es obtener información relevante a partir de una gran cantidad de información almacenada en una base de datos, pero de qué sirve esto si no sabemos qué es lo que necesitamos o estamos buscando, primeramente, necesitamos ubicar las necesidades y objetivos que tenemos que cumplir, la razón por la que necesitamos utilizar la minería de datos.

La minería de datos puede cumplir varios objetivos como es la toma de decisiones difíciles, así como de realizar estrategias a partir de datos o información relevante, comparación de resultados buscando la mejora, etc. El tener claro los objetivos que estamos buscando cumplir nos

facilitarán los siguientes pasos para realizar una buena minería de datos, debemos tomar en cuenta las reglas del negocio en las que se basa la compañía o entidad para la que vamos a realizar nuestra minería de datos, y alinear nuestros objetivos con los de la compañía.

Una vez determinado nuestro objetivo podemos utilizarlo para elegir nuestro modelo de minería de datos, como ya sabemos existen diferentes algoritmos para realizar minería de datos y cada uno de estos modelos refleja mejor ciertos tipos de resultados, es decir, no podemos elegir un algoritmo de minería de datos que nos dará como respuesta una fórmula matemática cuando lo que estamos buscando es una estrategia para mejorar nuestras ventas basadas en las tendencias de compra de las personas. (Los diferentes modelos de minería de datos y sus respectivos objetivos se explicaron previamente en esta guía).

3.1.1.2. Paso 2. Preparación de ambiente y selección de herramientas.

En el presente paso se va a revisar lo que se necesita para la creación de una base de datos ideal por lo que es necesario el conocer las diversas características de las herramientas como el formato en el que se deba encontrar la base de datos para el uso previo con el sistema de minería de datos; el cual es de suma importancia para conocer la herramienta y el tipo de almacenamiento adecuado con el que se desee trabajar o aplicar en las grandes cantidades de información en cualquier tipo de organización.

3.1.1.2.1. Preparación de ambiente

Para la preparación del ambiente se necesita la previa construcción de una base de datos, empleando cualquier tipo de herramienta, por lo que es necesario definir y diseñar claramente un modelo de base de datos en la herramienta que hayamos seleccionado para la construcción de la misma. Por lo que es fundamental tener conocimiento previo del como modelar una base de datos en función al problema que deseamos desarrollar, de tal forma que podamos

realizar registros en los campos predeterminados sin ningún tipo de dificultad.

Una vez ubicado el objetivo o el fin que deseamos obtener con la minería de datos, pasamos a la etapa donde necesitaremos conocer la información con la cual realizaremos la minería de datos. Es necesario saber qué tipo de información nos será útil al momento de realizar la minería de datos y que información es irrelevante con el objetivo de mejorar el resultado de la minería de datos e incluso acelerar el proceso, datos irrelevantes nos perjudicarán al momento de realizar nuestra minería ya que pueden generar información sin sentido o simplemente complicar la interpretación de resultados.

Muchas de las veces que se realiza minería de datos, la información y las bases de datos provienen de varios lugares y en diferentes formatos, totalmente ajenos los unos de los otros, por lo cual es necesario unir toda esta información en un almacenamiento que sea capaz de soportar los diferentes formatos en los que se encuentran la información o datos.

Es de gran importancia tomar en cuenta el modelo que se haya elegido para realizar nuestra minería de datos, ya que al momento de realizar nuestra minería de datos pueden existir conflictos entre los formatos de los datos con los que nos encontramos trabajando y con el método en sí, ya que el método no soporta tipos de datos que sean ajenos a su objetivo o simplemente el algoritmo con el que trabaja es incompatible.

Una vez determinada nuestra base de datos con la cual vamos a realizar nuestra minería de datos, procedemos a realizar la limpieza y corrección de los datos (como se vio anteriormente en la fase de preparación y limpieza de datos del proceso KDD). Hay que tener en cuenta que los datos de igual manera tienen que encontrase normalizados, esto con el fin de que no existan problemas en el momento de reunir la información o datos en un mismo ambiente capaz de controlar los formatos de la información.

Una vez que pasemos por nuestra clasificación de datos, así como el proceso de limpieza y estandarización deberíamos tener como

resultado nuestra superficie minable que será utilizada para la minería de datos.

3.1.1.2.2. Selección de una herramienta para minería de datos

Para seleccionar una herramienta de minería de datos se debe tomar en cuenta principalmente el objetivo que se desea cumplir o la meta que la organización desea alcanzar para obtener os mejores resultados posibles. Otros factores a considerar son las características de la herramienta como:

- Los tipos de datos: la compatibilidad de la herramienta con los diferentes tipos de datos, no siempre es la misma para todos los sistemas de minería de datos
- Funciones y metodologías: Los sistemas de minería de datos pueden ofrecer varias funciones, se deberá seleccionar el que posea las funciones necesarias, no el que posea más opciones.
- Herramientas de visualización: Las herramientas de minería ofrecen distintos maneras de visualizar los resultados, se deberá seleccionar aquel que permita una mejor representación del objetivo deseado.
- Interfaz de usuario: La facilidad de interacción del usuario con el sistema de minería.

3.1.1.2.2.1. Tipos de almacenamiento

Dependiendo de la herramienta de minería que se disponga a usar se maneja el almacenamiento de los datos y registros que se utilizaran pueden variar, es decir se deberá realizar la transformación apropiada de los registros, con el fin de que sean compatibles con la herramienta de minería. Estos formatos varían dependiendo de la herramienta, por ejemplo, en WEKA se trabaja con formatos .arff y en PENTAHO el formato .cvs

3.1.1.3. Paso 3. Empezando con la minería de datos.

Para empezar a realizar nuestro proceso de minería de datos debemos comenzar con la herramienta que utilizaremos, existen muchas herramientas disponibles para realizar minería de datos, tanto libres como licenciadas, debemos considerar una vez más nuestro objetivo a cumplir con la minería de datos y ver que herramienta podrá sernos de utilidad.

Debemos considerar la cantidad de atributos que utilizaremos en la metodología para minería de datos, ya que en muchos casos a pesar de utilizar el mismo modelo puede cambiar el resultado dependiendo de los atributos ingresados, así como de la cantidad de atributos, por lo que es necesario utilizar los diferentes modelos varias veces con diferentes variaciones de los atributos ingresados hasta encontrar el resultado más óptimo o que mejor se acople a nuestro objetivo. Un ejemplo de este tipo de casos es cuando se utiliza el modelo de minería de datos J48, entre más atributos se utilicen más preciso será el resultado, sin embargo, un exceso en la cantidad de atributos puede perjudicar el resultado si las variables no tienen sentido con el resultado buscado.

(Esto queda a la discreción de la persona que realizará la minería de datos, más adelante en esta guía se procederá con un ejemplo práctico de la minería de datos utilizando la herramienta de software libre WEKA).

3.1.1.4. Paso 4. Análisis e interpretación de resultados.

Una vez realizada la minería de datos procedemos a interpretar los resultados y las diferentes teorías que se puedan realizar a partir de los datos, muchas de las veces que se realiza minería de datos los resultados pueden encontrarse fuera de los datos esperados o representar inconsistencias, esto es debido a las variables y atributos que se utilizaron al momento de realizar la minería de datos, por lo que será necesario repetir el proceso de minería cambiando variables o incluso el método como tal, dependiendo de los resultados esperados. Si no existen complicaciones con respecto a las variables o el método es posible que

los datos nos reflejen algún problema existente con respecto a nuestra organización, empresa o fin con el que empezamos a realizar la minería, aquí es cuando tomamos los resultados e interpretamos lo que podría representar, para consecuentemente podamos tomar acciones o medidas en base a los resultados obtenidos.

3.2. MOTIVOS PARA LA EXTRACCIÓN Y ANÁLISIS DE LA INFORMACIÓN

A partir de diversas experiencias las organizaciones optan por seguir alguna metodología, en donde se encuentre definido los procedimientos a seguir para generar calidad en sus resultados forjando casos exitosos o mejores soluciones.

3.2.1. Comprensión del negocio

En este punto la organización o el individuo debe visualizar o establecer los requerimientos y objetivos del negocio, por lo que es necesario que el mismo realice una evaluación sobre la situación en la que se encuentra dicha organización.

Una vez finalizado el proceso de evaluación se procede a establecer los principales objetivos para lo que se va a emplear la minería de datos y así generando un plan del proyecto en donde estará compuesta de las herramientas que se van a emplear, las técnicas a usar y el equipo en donde se van a ejecutar.

3.2.2. Comprensión de los datos

Es necesario tener en cuenta los objetivos a los que está enfocada la organización, considerando como principales características a la recopilación de todos los datos iniciales, a la descripción detallada de

los datos, a la exploración minuciosa de los datos y finalmente a la verificación sobre la validez y calidad de los datos obtenidos.

3.2.3. Preparación de los datos

Debemos considerar el conjunto de datos en donde cada una de sus variables contienen sus valores o datos predeterminados, los cuales dichos valores pueden ser empleados para: una selección de datos, se logra realizar limpieza de datos, se efectúa la construcción de datos, los datos pueden ser integrados y datos pueden ser formateados.

3.2.4. Modelado

En el modelamiento de los datos se realiza la aplicación de las diversas metodologías o técnicas para el minado de datos, en los cuales es necesario seleccionar la técnica de modelado, realizar un diseño de la evaluación, crear o construir el modelo y finalmente evaluar el modelo creado.

3.2.5. Evaluación

En esta fase se determina la validez de las fases anteriores en donde se ve que todo lo realizado es útil y se integra a las necesidades de la organización. Para proceder con la verificación en esta fase es necesario realizar una evaluación de todos los resultados obtenidos, revisar minuciosamente los procesos y posteriormente establecer los pasos a seguir o detallar las acciones que se deban tomar.

3.2.6. Despliegue

Visualizar la importancia de los modelos efectuados y ver el nivel de influencia o certeza que puedan tener los mismos al momento de integrase con tareas específicas para la toma de decisiones.

Es importante planificar el despliegue, para ejecutar un plan de monitoreo y mantenimiento constante, generando de esto un informe detallado final, con el cual se realizará una revisión total al proyecto.

3.2.7. Forma de como presentar los hallazgos

La carta de presentación de las diversas organizaciones se compone del procedimiento de la obtención de resultados de calidad los cuales fueron conseguidos por medio de alguna metodología o algoritmo. Por lo que se propone la creación de una guía o un informe detallado en donde se pueda comprender los resultados obtenido generando así una un documento que sea de fácil interpretación y entendible para el usuario final.

Este informe nos permitirá extraer diversas ideas para nuevas soluciones a los problemas de los proyectos de una organización

4. CAPÍTULO 4: VALIDACIÓN DE LA GUÍA METODOLÓGICA CON UN EJEMPLO PRÁCTICO

En este capítulo se presenta dos diferentes aplicaciones prácticas de la Guía Metodológica sobre Minería de Datos, tanto en WEKA como en Pentaho.

EJEMPLO PRÁCTICO UTILIZANDO LA HERRAMIENTA WEKA

4.1. Paso 1. Identificar las necesidades

Utilizaremos la herramienta WEKA con el fin de realizar una práctica sencilla y entendible, debido a la facilidad de uso que nos brinda la interfaz gráfica de esta herramienta.

En este ejemplo buscamos obtener información de una base de datos a cerca de pruebas de diabetes realizadas a pacientes mujeres que viven en el estado de Arizona (USA), con el fin de buscar anomalías en el proceso de pruebas, así como la competitividad de los doctores y enfermeras a cargo de realizar dicho examen.

Como algoritmo de minería de datos utilizaremos la generación de un árbol de decisión con el algoritmo J48[29], esto también por conveniencia de las variables que se encuentran todas ellas en formato numérico y al hecho de que este algoritmo es compatible con dichas variables.

4.2. Paso 2. Preparación de ambiente y selección de herramientas

4.2.1. Preparación de ambiente

Es necesario conocer el propósito del desarrollo de una base de datos en el cual estará representada la información que se va a utilizar, tomando en cuenta que se almacenará grandes cantidades de información; por lo

[29] Algoritmo J48: Algoritmo de minería de datos basado en la generación de árboles de decisión. Utilizado para la clasificación y generación de hipótesis.

que es importante considerar la calidad y limpieza de los datos que debería existir en cada uno de los campos para así generar una superficie minable óptima.

Para la creación de nuestra superficie minable se deberá discernir entre toda la información disponible, la de mayor importancia o relevancia para cumplir el objetivo que se busca con mejores resultados, por ejemplo no podemos considerar la información de una base de datos a cerca de enfermedades si nuestro objetivo es el de conseguir las tendencias de venta en ropa, debe existir una relación coherente entre la información obtenida. Un buen consejo es conocer la fiabilidad de las fuentes de la que proviene la información, esto ayudara a mejorar la calidad de los datos ya que una fuente confiable contendrá menos “ruido” en sus datos.

Sin embargo ninguna fuente de información se encuentra ausente de ruido, por lo que es necesario corregirla mediante un proceso de limpieza. (Proceso explicado previamente en la sección 2.4.1.1 Etapas y fases que conforman el proceso KDD), esto conseguirá una superficie minable lista para el proceso de minería.

4.2.2. Selección de una herramienta para minería de datos

En este paso procedemos a inicializar WEKA que es nuestra herramienta de software. Como vemos en la Figura 4.2.1.1.

Figura 4.2.1.1. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

Podemos apreciar las diferentes opciones a las cuales podemos acceder utilizando esta herramienta.

Procedemos a utilizar la opción “Explorer” para desplegar la ventana que nos permitirá ingresar la información de la base de datos que utilizaremos para realizar la minería de datos. Que vemos en la Figura 4.2.1.2.

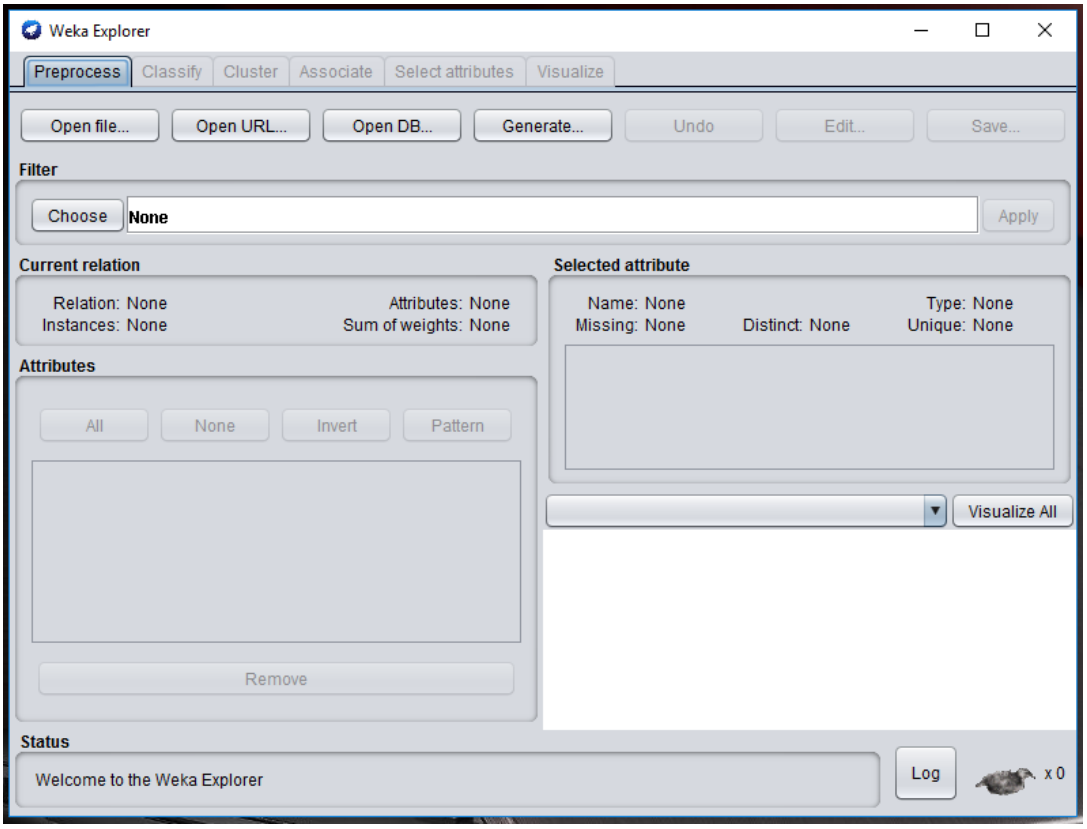


Figura 4.2.1.2. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

Se desplegará la siguiente pantalla donde podemos apreciar que existen varias pestañas, las cuales nos servirán para realizar nuestra minería de datos, así como analizar los resultados que obtendremos al final. (Durante el ejemplo profundizaremos en el uso de dichas pestañas)

- **Preprocess:** Esta pestaña nos permitirá cargar nuestra base de datos, así como una breve gráfica de los datos, se pueden utilizar filtros de tal manera que podremos escoger los atributos que deseamos utilizar para la minería de datos.

- **Classify:** En esta pestaña podremos escoger con que algoritmo de minería de datos vamos a trabajar sobre nuestra base de datos. Por ejemplo, el algoritmo J48, (que utilizaremos en este ejemplo), el algoritmo de Bayes, etc.
- **Cluster:** Nos permitirá realizar una minería de datos más precisa ofreciéndonos resultados estadísticos, porcentajes de errores.
- **Associate:** Nos permite conocer las relaciones existentes entre los datos.
- **Select attributes:** Aquí podremos seleccionar los atributos más importantes de nuestra base de datos, los cuales tendrán más relevancia al momento de realizar nuestro proceso de minería.
- **Visualize:** Nos presenta graficas resultantes de nuestro proceso de minería.

A continuación procederemos a cargar nuestra base de datos (en el formato .arff) mediante la pestaña “Preprocess”, seleccionamos “Open File”, un browser aparecerá y procederemos a cargar nuestra base de datos. Como vemos en la Figura 4.2.1.3.

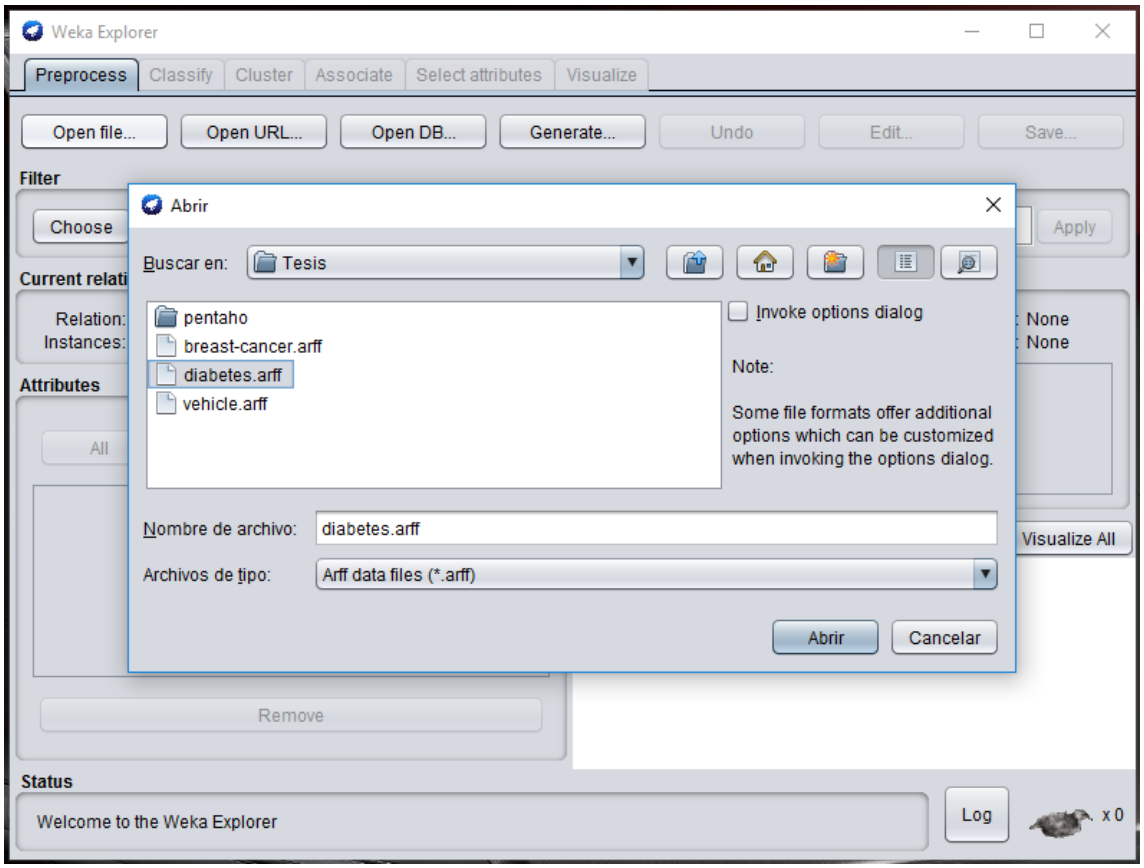


Figura 4.2.1.3. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

Una vez cargada nuestra base de datos podremos ver las gráficas de los datos presentes en la base de datos. Que se presenta en la Figura 4.2.1.4.

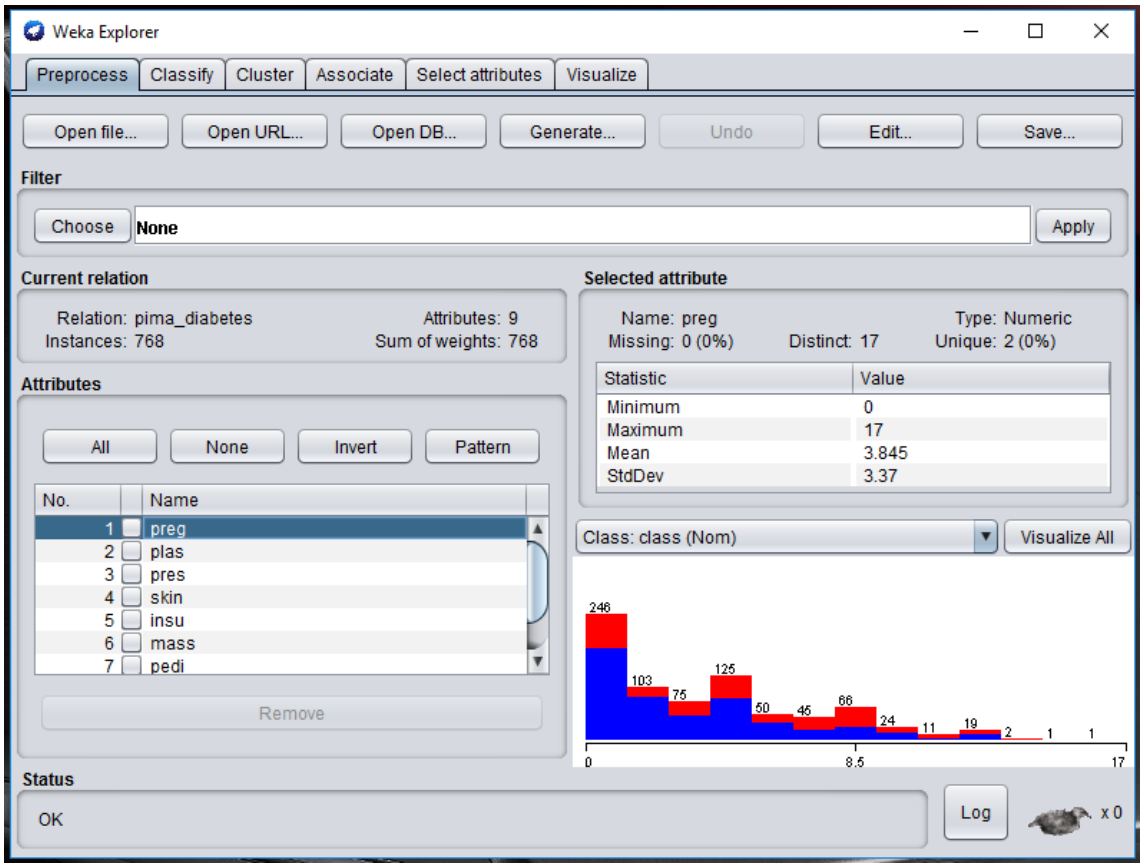


Figura 4.2.1.4. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

La grafica variará dependiendo del atributo que deseemos inspeccionar. Como se presenta en la Figura 4.2.1.5.

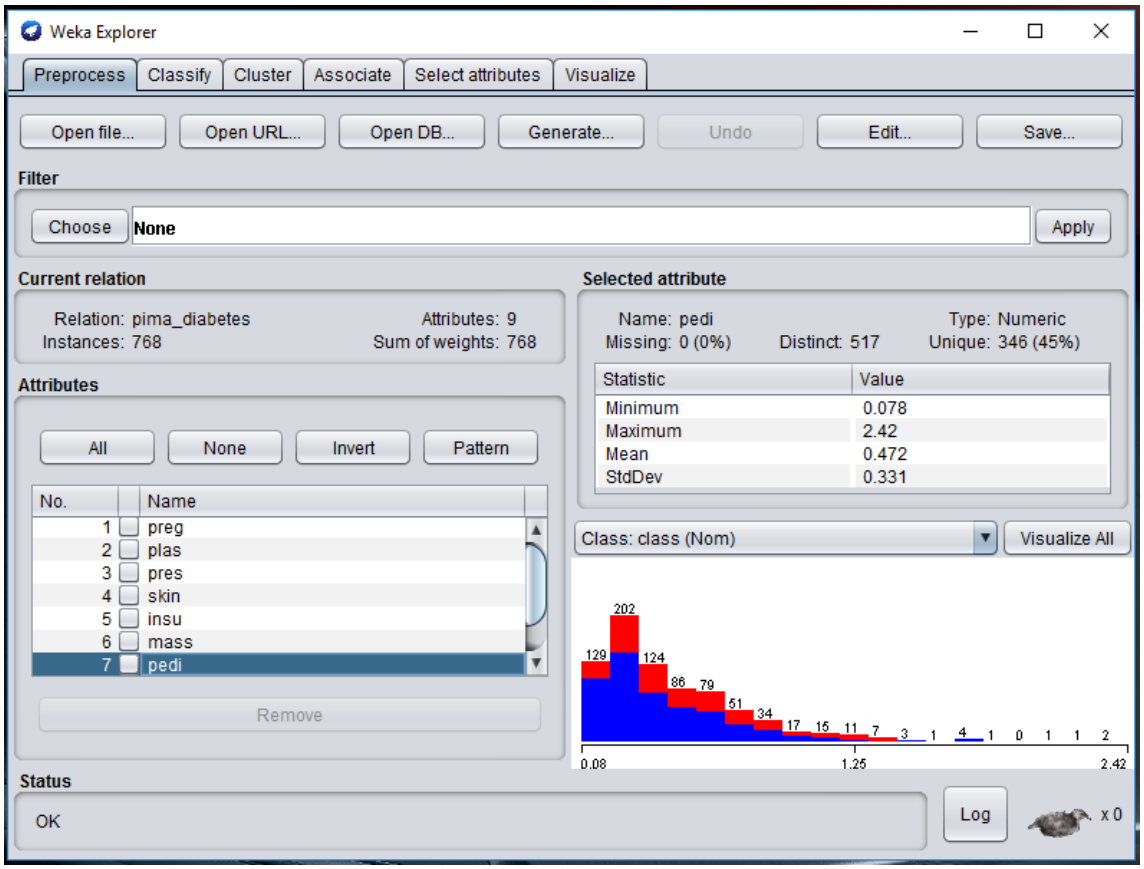
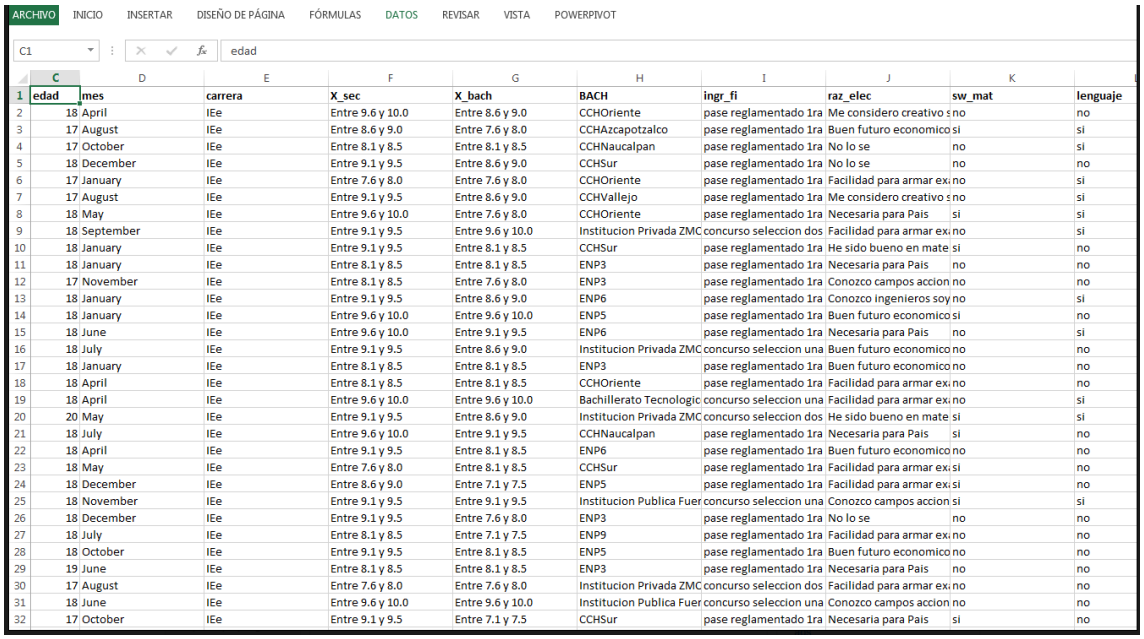


Figura 4.2.1.5. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

4.2.2.1. Tipos de almacenamiento

En primer lugar necesitaremos nuestra base de datos (superficie minable) en el formato .arff que es una extensión de archivo con la que trabaja la herramienta WEKA.

Procedemos a exportar la base de datos (superficie minable) a Excel como en la Figura 4.2.2.1.1. y procedemos a poner entre comillas simples los campos que contengan texto o espacios en blanco, y lo guardamos como archivo con la extensión .csv.(MS-DOS).



ARCHIVO INICIO INSERTAR DISEÑO DE PÁGINA FÓRMULAS DATOS REVISAR VISTA POWERPIVOT

C1 edad

	C	D	E	F	G	H	I	J	K	L
1	edad	mes	carrera	X_sec	X_bach	BACH	ingr_fi	raz_elec	sw_mat	lenguaje
2	18	April	IEe	Entre 9.6 y 10.0	Entre 8.6 y 9.0	CCHOriente	pase reglamentado 1ra	Me considero creativo s	no	no
3	17	August	IEe	Entre 8.6 y 9.0	Entre 7.6 y 8.0	CCHAzcapotzalco	pase reglamentado 1ra	Buen futuro economico	si	si
4	17	October	IEe	Entre 8.1 y 8.5	Entre 8.1 y 8.5	CCHNaucalpan	pase reglamentado 1ra	No lo se	no	si
5	18	December	IEe	Entre 9.1 y 9.5	Entre 8.6 y 9.0	CCHSur	pase reglamentado 1ra	No lo se	no	no
6	17	January	IEe	Entre 7.6 y 8.0	Entre 7.6 y 8.0	CCHOriente	pase reglamentado 1ra	Facilidad para armar ex	no	si
7	17	August	IEe	Entre 9.1 y 9.5	Entre 8.6 y 9.0	CCHVallejo	pase reglamentado 1ra	Me considero creativo s	no	si
8	18	May	IEe	Entre 9.6 y 10.0	Entre 7.6 y 8.0	CCHOriente	pase reglamentado 1ra	Necesaria para Pais	si	si
9	18	September	IEe	Entre 9.1 y 9.5	Entre 9.6 y 10.0	Institucion Privada ZMC	concurso seleccion dos	Facilidad para armar ex	no	si
10	18	January	IEe	Entre 9.1 y 9.5	Entre 8.1 y 8.5	CCHSur	pase reglamentado 1ra	He sido bueno en mate	si	no
11	18	January	IEe	Entre 8.1 y 8.5	Entre 8.1 y 8.5	ENP3	pase reglamentado 1ra	Necesaria para Pais	no	no
12	17	November	IEe	Entre 8.1 y 8.5	Entre 7.6 y 8.0	ENP3	pase reglamentado 1ra	Conozco campos accion	no	no
13	18	January	IEe	Entre 9.1 y 9.5	Entre 8.6 y 9.0	ENP6	pase reglamentado 1ra	Conozco ingenieros soy	no	si
14	18	January	IEe	Entre 9.6 y 10.0	Entre 9.6 y 10.0	ENP5	pase reglamentado 1ra	Buen futuro economico	si	no
15	18	June	IEe	Entre 9.6 y 10.0	Entre 9.1 y 9.5	ENP6	pase reglamentado 1ra	Necesaria para Pais	no	si
16	18	July	IEe	Entre 9.1 y 9.5	Entre 8.6 y 9.0	Institucion Privada ZMC	concurso seleccion una	Buen futuro economico	no	no
17	18	January	IEe	Entre 8.1 y 8.5	Entre 8.1 y 8.5	ENP3	pase reglamentado 1ra	Buen futuro economico	no	no
18	18	April	IEe	Entre 8.1 y 8.5	Entre 8.1 y 8.5	CCHOriente	pase reglamentado 1ra	Facilidad para armar ex	no	no
19	18	April	IEe	Entre 9.6 y 10.0	Entre 9.6 y 10.0	Bachillerato Tecnologic	concurso seleccion una	Facilidad para armar ex	no	si
20	20	May	IEe	Entre 9.1 y 9.5	Entre 8.6 y 9.0	Institucion Privada ZMC	concurso seleccion dos	He sido bueno en mate	si	si
21	18	July	IEe	Entre 9.6 y 10.0	Entre 9.1 y 9.5	CCHNaucalpan	pase reglamentado 1ra	Necesaria para Pais	si	no
22	18	April	IEe	Entre 9.1 y 9.5	Entre 8.1 y 8.5	ENP6	pase reglamentado 1ra	Buen futuro economico	no	no
23	18	May	IEe	Entre 7.6 y 8.0	Entre 8.1 y 8.5	CCHSur	pase reglamentado 1ra	Facilidad para armar ex	si	no
24	18	December	IEe	Entre 8.6 y 9.0	Entre 7.1 y 7.5	ENP5	pase reglamentado 1ra	Facilidad para armar ex	si	no
25	18	November	IEe	Entre 9.1 y 9.5	Entre 9.1 y 9.5	Institucion Publica Fuer	concurso seleccion una	Conozco campos accion	si	si
26	18	December	IEe	Entre 9.1 y 9.5	Entre 7.6 y 8.0	ENP3	pase reglamentado 1ra	No lo se	no	no
27	18	July	IEe	Entre 8.1 y 8.5	Entre 7.1 y 7.5	ENP9	pase reglamentado 1ra	Facilidad para armar ex	no	no
28	18	October	IEe	Entre 9.1 y 9.5	Entre 8.1 y 8.5	ENP5	pase reglamentado 1ra	Buen futuro economico	no	no
29	19	June	IEe	Entre 8.1 y 8.5	Entre 8.1 y 8.5	ENP3	pase reglamentado 1ra	Necesaria para Pais	no	no
30	17	August	IEe	Entre 7.6 y 8.0	Entre 7.6 y 8.0	Institucion Privada ZMC	concurso seleccion dos	Facilidad para armar ex	no	no
31	18	June	IEe	Entre 9.6 y 10.0	Entre 9.6 y 10.0	Institucion Publica Fuer	concurso seleccion una	Conozco campos accion	no	no
32	17	October	IEe	Entre 9.1 y 9.5	Entre 7.1 y 7.5	CCHSur	pase reglamentado 1ra	Necesaria para Pais	si	no

Figura 4.2.2.1.1. Paso 2, Elaborado por: (Anónimo, 2014)

Procedemos a modificar este archivo mediante un bloc de notas como poniéndolo en formato .arff, para esto procedemos a colocar @Relation precediendo el nombre de nuestra base de datos y @Attribute precediendo los nombres de los campos de la base de datos seguido del tipo de dato.

Por último, colocamos @Data antes de toda la información de la base de datos.

Como podemos observar en la Figura 4.2.2.1.2. y Figura 4.2.2.1.3.

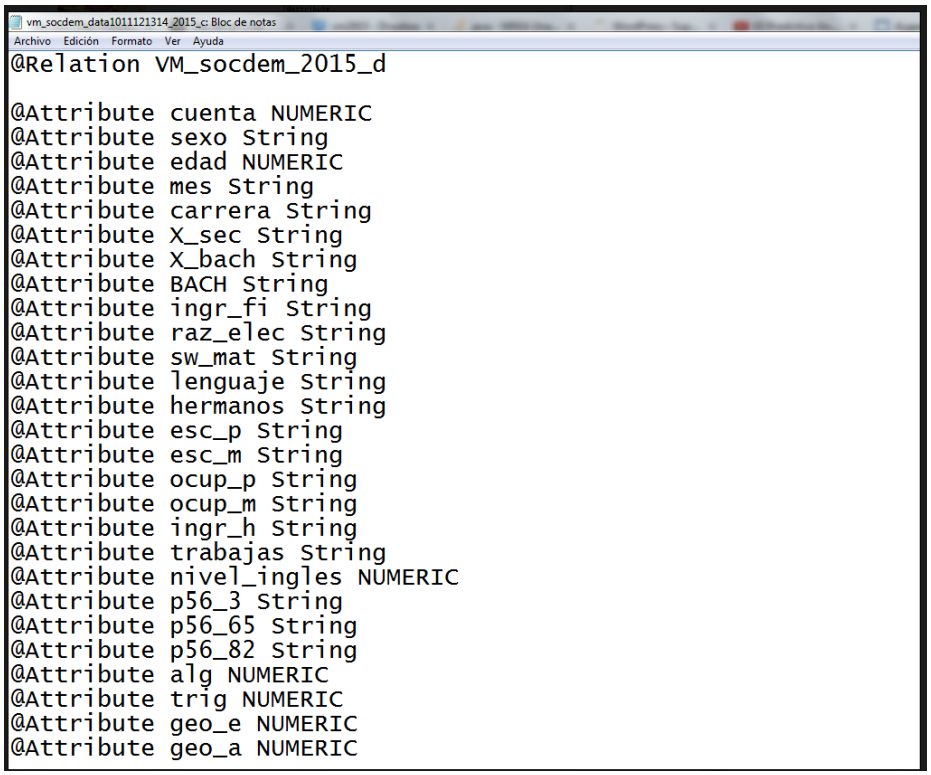


Figura 4.2.2.1.2. Paso 2, Elaborado por: (Anónimo, 2014)

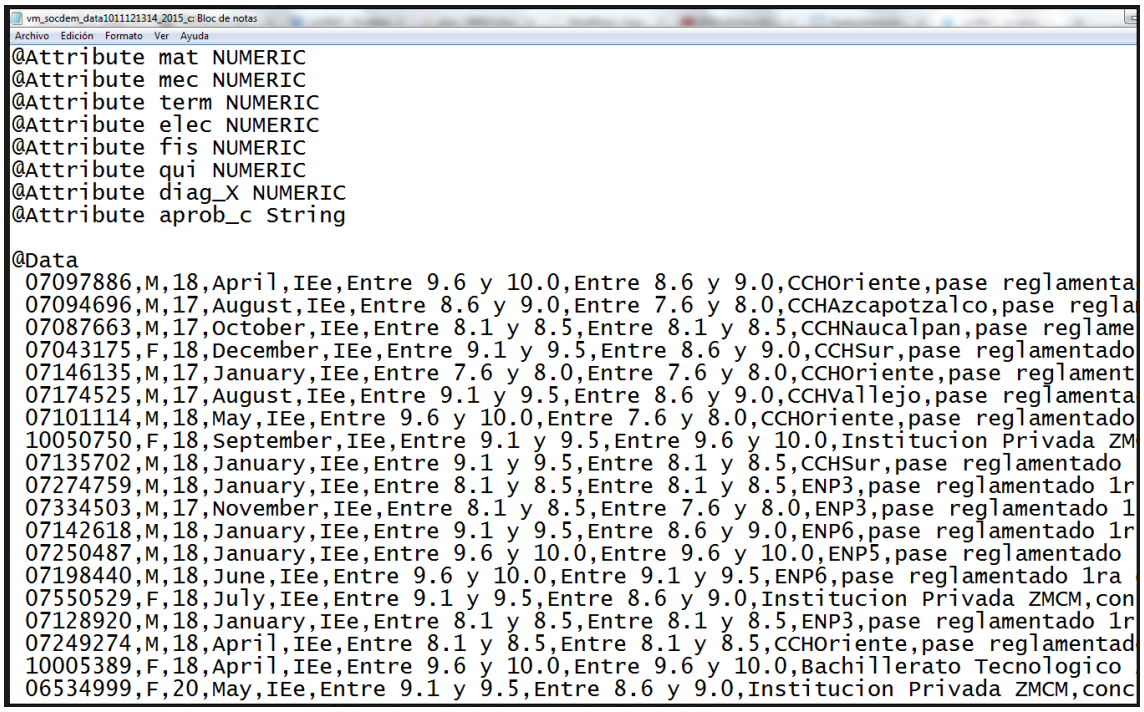


Figura 4.2.2.1.3. Paso 2, Elaborado por: (Anónimo, 2014)

Finalmente guardamos el archivo con todas las modificaciones con la extensión .arff

4.3. Paso 3. Empezando con la minería de datos.

Una vez cargada nuestra base de datos seremos capaces de utilizar las demás pestañas que nos ofrece la herramienta para nuestra minería de datos.

En esta pantalla podemos realizar el filtrado de atributos que deseemos utilizar en nuestra minería, mediante el botón “Choose”, si así lo deseamos, también es posible eliminar los atributos de la base de datos que no creamos convenientes, simplemente seleccionamos el atributo que deseamos eliminar y presionamos el botón “Remove”. Como se ve en la Figura 4.3.1.

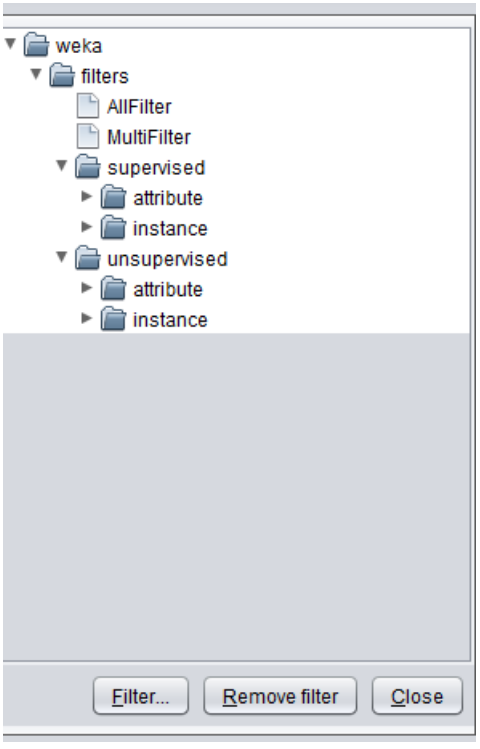


Figura 4.3.1. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Los filtros a aplicarse pueden ser supervisados o no supervisados, eligiendo si trabajar bajo los atributos o las instancias. Dependerá de

nuestro objetivo deseado el si es necesario aplicarse algún filtro, en el caso de este ejemplo, no se aplicará ningún filtro.

Una vez cargada nuestra base de datos procederemos a escoger nuestro algoritmo de minería de datos, el cual aplicaremos en nuestra base de datos.

En la pestaña “Classify” podemos observar la opción “Choose”, es aquí donde determinaremos que tipo de algoritmo aplicaremos. Lo mencionado anteriormente se puede ver en la Figura 4.3.2.

Figura 4.3.2. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Una vez más dependerá de que tipo de resultado estemos buscando, el que determinará que algoritmo será el necesario. Para este ejemplo utilizaremos el algoritmo en base a árboles de decisión, que será el J48.

Una vez seleccionado nuestro algoritmo se activarán varias opciones en el recuadro de Test Options, con las cuales podemos controlar el tipo de entrenamiento y aprendizaje que WEKA utilizara al momento de realizar la minería. Como se presenta en la Figura 4.3.3.

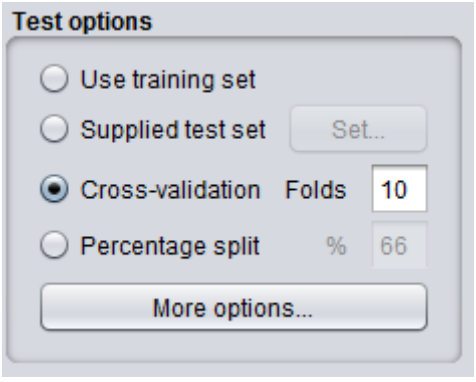


Figura 4.3.3. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Use training set: WEKA utilizara los archivos ya existentes, así como registros como base para realizar la minería de datos

Supplied test set: Se puede seleccionar un archivo de extensión arff, el cual se encargará del aprendizaje de la minería de datos.

Cross-validation: El aprendizaje se realizará a partir de una cierta cantidad de campos o atributos que se requiera.

Percentage Split: Se puede determinar el porcentaje de datos que será utilizado durante el aprendizaje, el cual se aplicará posteriormente al porcentaje de datos restantes para realizar la minería.

Para nuestro ejemplo seleccionamos la opción de “Cross-validation”, a continuación, podemos seleccionar el atributo base en el cual se realizará la minería de datos que se presenta en la Figura 4.3.4., por defecto siempre se selecciona el último atributo de la base de datos, pero se puede seleccionar cualquier atributo, en nuestro caso dejaremos el atributo por defecto. Para interpretación significa que realizaremos una predicción de si el examen de diabetes fue positivo o no.

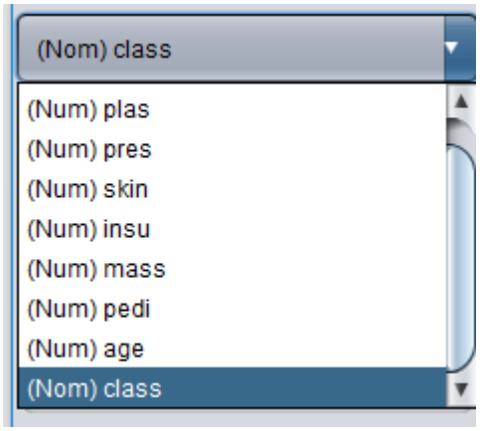


Figura 4.3.4. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Una vez seleccionado el algoritmo con el que vamos a trabajar, presionamos “Start” para efectuar la minería de datos

Los resultados aparecerán en el recuadro a lado derecho. Como se visualiza en la Figura 4.3.4.

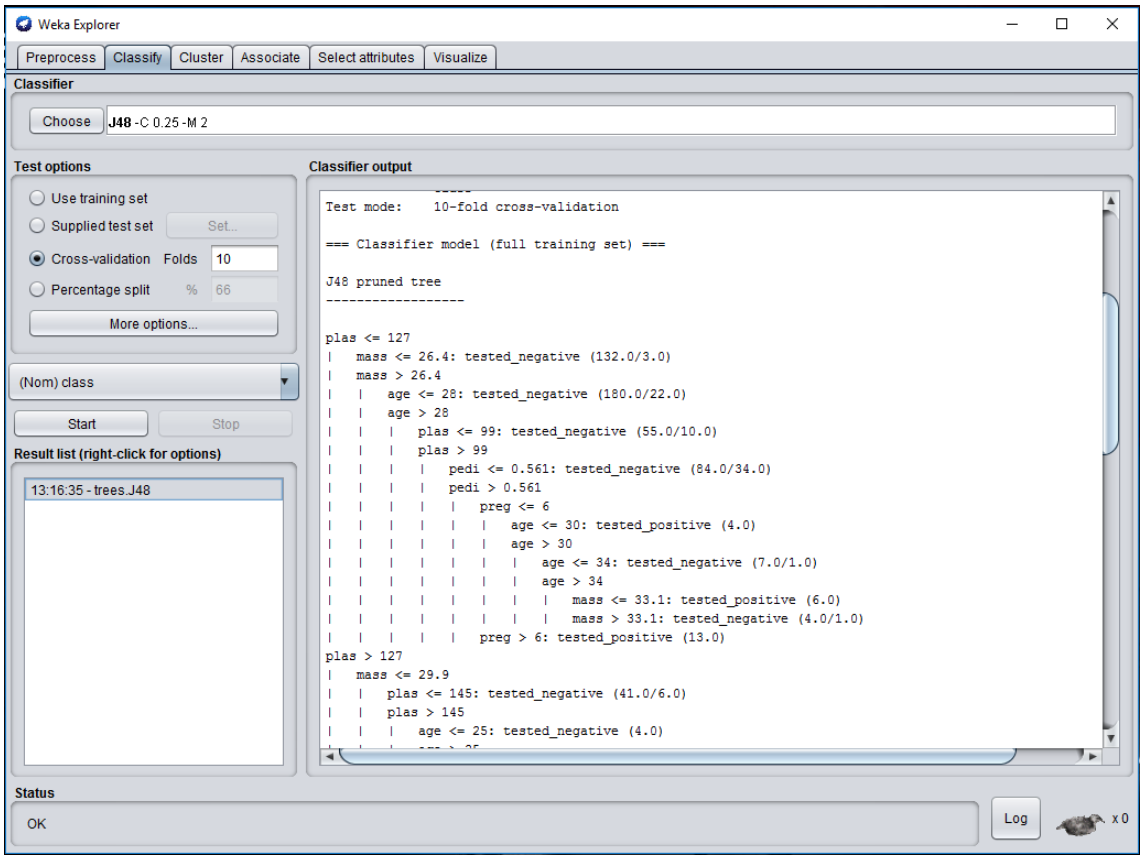


Figura 4.3.4. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Podemos apreciar la información del árbol de decisiones que se generó en la Figura 4.3.5., así como el número de hojas, el tamaño del árbol, el número de instancias y porcentajes de error al generar el árbol.

```
plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561
|   |   |   |   |   preg <= 6
|   |   |   |   |   |   age <= 30: tested_positive (4.0)
|   |   |   |   |   |   age > 30
|   |   |   |   |   |   |   age <= 34: tested_negative (7.0/1.0)
|   |   |   |   |   |   |   age > 34
|   |   |   |   |   |   |   |   mass <= 33.1: tested_positive (6.0)
|   |   |   |   |   |   |   |   mass > 33.1: tested_negative (4.0/1.0)
|   |   |   |   |   preg > 6: tested_positive (13.0)
```

Figura 4.3.5. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Se puede interpretar en la Figura 4.3.6. y en la Figura 4.3.7. que de las personas que presentan una concentración de glucosa inferior o igual a 127, aquellas que tienen un índice de masa corporal inferior o igual a 26.4, 3 personas de 132 presentan un examen de diabetes negativo. En caso de que su índice de masa muscular sea superior es necesario considerar su edad, si es menor a los 28 años, 22 personas de un grupo de 180 presentan un examen de diabetes negativo. El árbol sigue expandiéndose hasta contemplar todos los parámetros ingresados.

```
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pres <= 82
|   |   |   |   |   |   |   pedi <= 0.396: tested_positive (8.0/1.0)
|   |   |   |   |   |   |   pedi > 0.396: tested_negative (3.0)
|   |   |   |   |   |   pres > 82: tested_negative (4.0)
|   |   |   |   age > 61: tested_negative (4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pres <= 61: tested_positive (15.0/1.0)
|   |   |   pres > 61
|   |   |   |   age <= 30: tested_negative (40.0/13.0)
|   |   |   |   age > 30: tested_positive (60.0/17.0)
|   |   plas > 157: tested_positive (92.0/12.0)
```

Figura 4.3.6. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

```
Number of Leaves  :     20

Size of the tree :      39


Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         567               73.8281 %
Incorrectly Classified Instances       201               26.1719 %
Kappa statistic                          0.4164
Mean absolute error                      0.3158
Root mean squared error                  0.4463
Relative absolute error                 69.4841 %
Root relative squared error             93.6293 %
Total Number of Instances              768
```

Figura 4.3.7. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Es necesario tomar en cuenta la matriz de confusión que nos presenta el programa, ya que aquí es donde podemos ver el grado de confiabilidad de los resultados obtenidos. Esto se puede interpretar de la siguiente forma. Los datos que se encuentran en la diagonal principal de la matriz deben ser mayores que los datos en sus respectivas columnas. Como se presenta en la Figura 4.3.8.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 407  93 |   a = tested_negative
 108 160 |   b = tested_positive
```

Figura 4.3.8. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

En el caso de nuestro ejemplo, 407 > 108 y 160 > 93 por lo tanto los resultados son confiables. La matriz de confusión variará dependiendo del atributo base que se eligió previamente.

Si así lo deseamos podemos visualizar el árbol de decisión presentado en la Figura 4.3.9., realizando clic derecho en la lista de resultados, “Visualizae tree”.

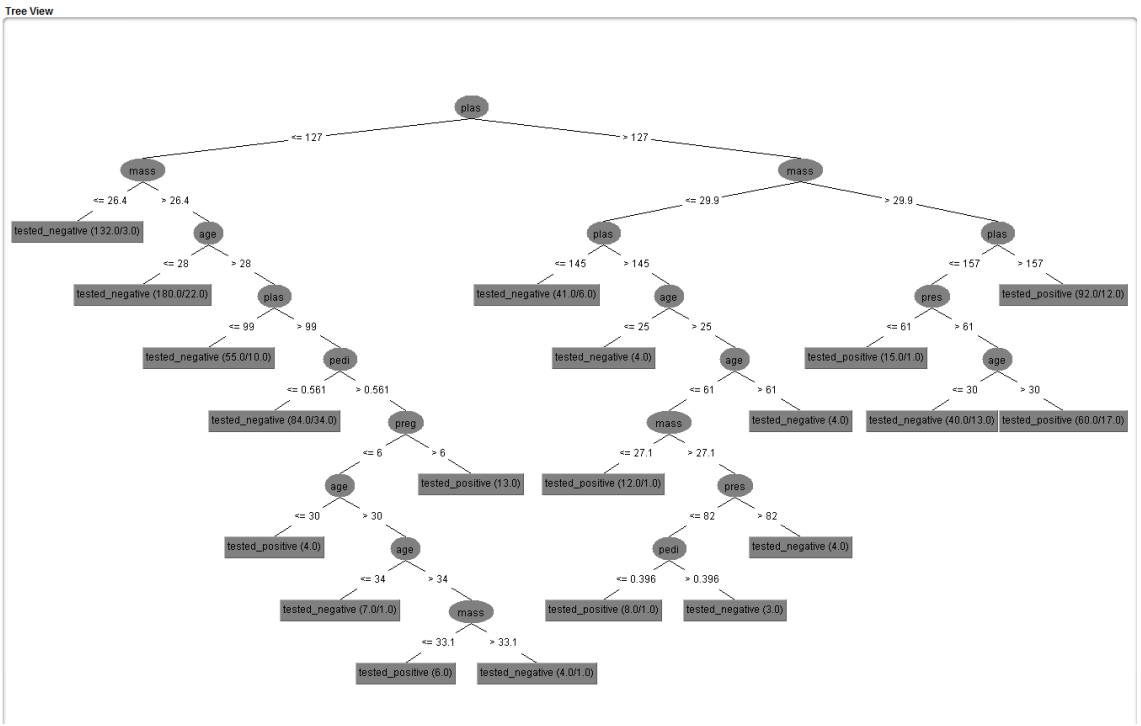


Figura 4.3.9. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Si analizamos los resultados arrojados por la herramienta podemos observar que existe una precisión de 73.83%, donde de 500 pacientes que tenían resultado de examen de diabetes negativo, los resultados indican que 93 podrían tener un examen positivo. Del otro lado, de 268 resultados positivos que se encontraban en la base de datos, los resultados indican que 108 pacientes deberían tener un resultado de examen negativo.

Si así lo deseamos podemos realizar un cluster que nos permitirá realizar varias simulaciones de minería de datos, con diferentes cantidades de datos una de la otra, lo que nos permitirá ver las diferentes variaciones en los resultados dependiendo de la sección de datos que se analice.

Para esto procedemos a la pestaña “Cluster”, seleccionamos el algoritmo con el cual realizaremos nuestro cluster, en nuestro caso utilizaremos SimpleKMeans, un algoritmo muy eficaz y preciso. Presionamos “Start” y los resultados se desplegarán del lado derecho. Como se observa en la Figura 4.3.10.

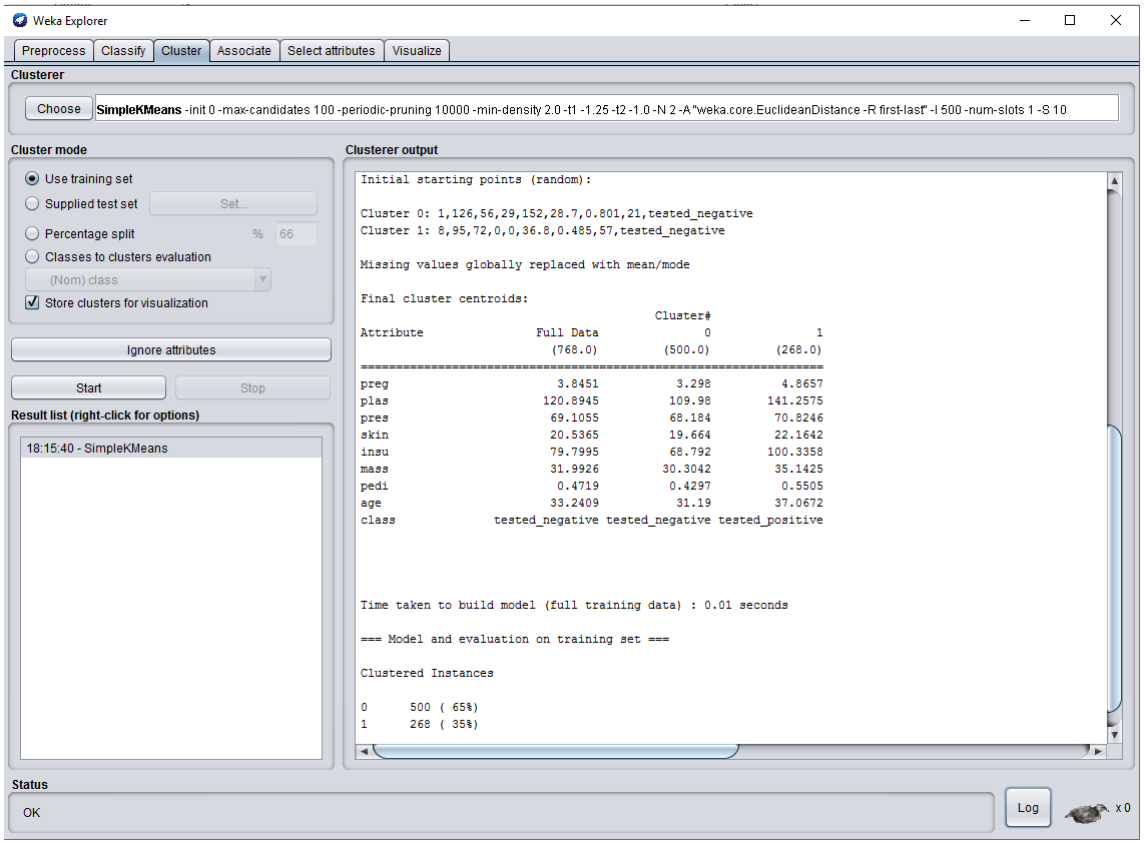


Figura 4.3.10. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Podemos ver en los resultados en la Figura 4.3.11. que se realizaron 4 iteraciones en los 3 diferentes clusters, en el primer cluster podemos ver que se ha utilizado todos los datos de la base, mientras en el cluster #0 se usaron 500 registros y en el cluster #1 los 268 registros restantes. Además de la cantidad de registros usados en los clusters, también nos muestra los promedios de cada uno de los atributos que se encuentran en la base de datos.

En el primer cluster podemos ver que la mayoría de pacientes presentan un examen negativo, al igual que el cluster #0, mientras que en el cluster #1 la mayoría presenta un resultado positivo en su examen.

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 149.5177664581119

Initial starting points (random):

Cluster 0: 1,126,56,29,152,28.7,0.801,21,tested_negative
Cluster 1: 8,95,72,0,0,36.8,0.485,57,tested_negative

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0               1
                          (768.0)         (500.0)         (268.0)
=============================================================================
preg                       3.8451           3.298          4.8657
plas                     120.8945          109.98        141.2575
pres                      69.1055          68.184         70.8246
skin                      20.5365          19.664         22.1642
insu                      79.7995          68.792        100.3358
mass                      31.9926         30.3042         35.1425
pedi                       0.4719          0.4297          0.5505
age                       33.2409           31.19         37.0672
class              tested_negative tested_negative tested_positive



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      500 ( 65%)
1      268 ( 35%)
```

Figura 4.3.11. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Otra de las opciones que nos presenta esta herramienta es el conocer cual o cuales de los atributos que se encuentran en la base de datos son los más importantes o determinantes al momento de realizar la minería. Para ver esta información procedemos a la pestaña de “Select Attributes”. Esto se puede observar en la Figura 4.3.12.

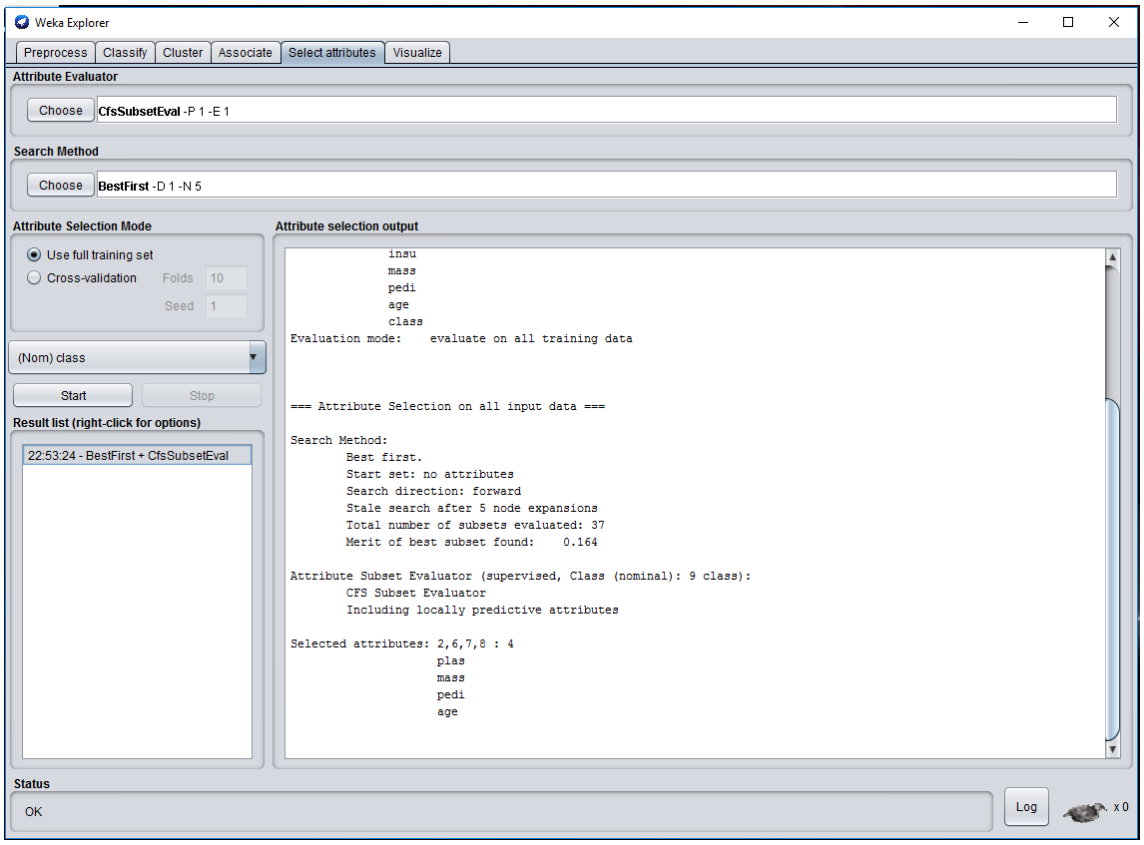


Figura 4.3.12. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

En nuestro caso podemos apreciar que los atributos más importantes son la concentración de glucosa, el índice de masa corporal, la edad y la función de diabetes.

Y finalmente la última herramienta que nos ofrece WEKA es la visualización, que nos permitirá apreciar gráficas de los atributos y sus datos, permitiéndonos conocer las correlaciones existentes entre los atributos.

Para acceder a esta opción simplemente seleccionamos la pestaña “Visualize”; como se presenta a en la Figura 4.3.13. y en la Figura 4.3.14.

Figura 4.3.13. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

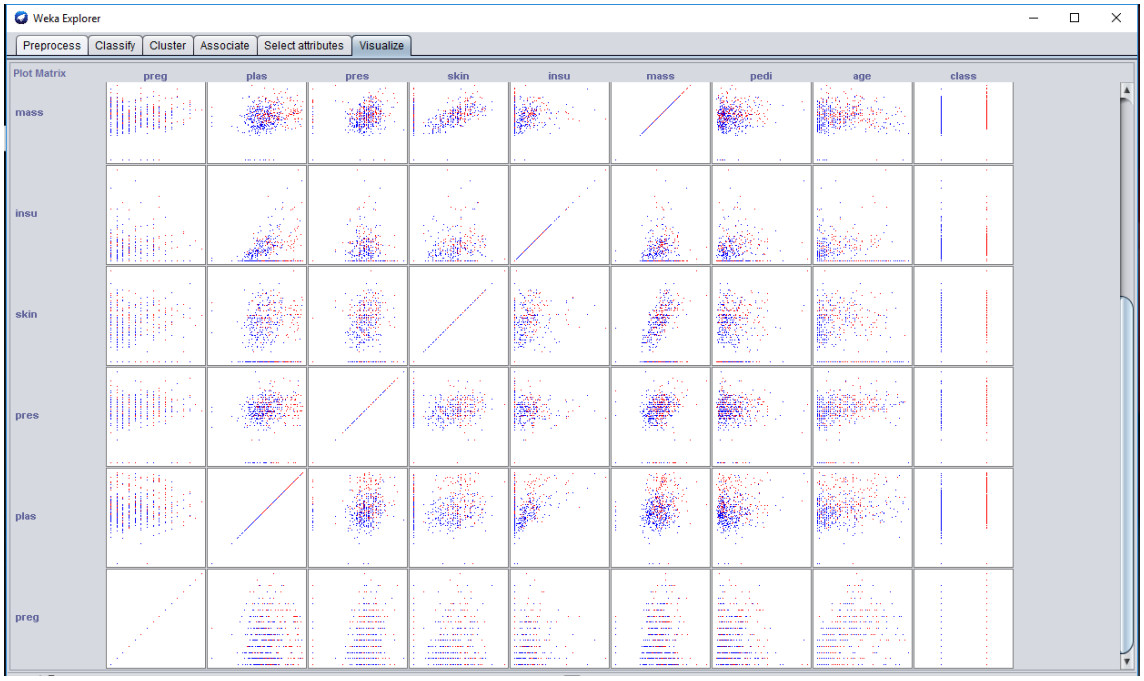


Figura 4.3.14. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Si interpretamos estas gráficas podemos darnos cuenta que en nuestro caso el número de embarazos de las pacientes no tienen ninguna relación con los demás atributos, es decir que no influye su edad, peso u otro

factor. De igual manera se puede apreciar que los resultados de los exámenes no están directamente relacionados con ningún otro atributo, en otras palabras, el hecho de que haya estado embarazada o no, o que haya superado cierto rango de edad, no influirá directamente con los resultados.

4.4. Paso 4. Análisis e interpretación de resultados.

Ahora que hemos analizado todas las opciones que nos ofrece WEKA, así como todos los resultados y análisis podemos proseguir a interpretar que significan.

Si nos basamos en los primeros resultados que nos ofrece la herramienta a través de “Classify”, podemos ver que existen anomalías en cuanto a los resultados de los exámenes, ya que los datos nos han mostrado la posibilidad de que ciertos pacientes hayan recibido diagnósticos erróneos, podemos comprobar la confiabilidad del resultado con la matriz de confusión.

Mediante los resultados del clusterizado podemos distinguir que diferentes grupos de datos han mostrado comportamientos diferentes, como en el cluster #1, donde la tendencia es que las pacientes presenten un resultado positivo, si prestamos mayor atención a este grupo de datos, podemos apreciar que los promedios de concentración de glucosa, la presión sanguínea y la concentración de insulina son superiores en comparación a los otros dos grupos de datos, donde el promedio es menor, lo cual nos puede indicar que estos atributos son los más determinantes en los resultados de los exámenes. Esto como nuestra apreciación al momento de ver los datos, se puede comprobar con los resultados que podemos apreciar mediante la pestaña “Select Attributes”, en la cual nos indica que los atributos más importantes son la concentración de glucosa, el índice de masa corporal, la edad y la función de diabetes.

Una vez más la concentración de glucosa aparece como uno de los atributos con importancia al momento de realizar un examen. Con esto podemos corroborar que realmente es un atributo influyente en el proceso

Y finalmente las gráficas de los datos nos muestran las relaciones entre los atributos, si somos más minuciosos al momento de observar las gráficas, podemos darnos cuenta de que una serie de datos, provenientes de exámenes con resultado negativo, se concentran en la gráfica de concentración de glucosa vs la función de diabetes, indicando una posible relación entre estos dos atributos, este tipo de concentración también se puede observar en la gráfica de concentración de glucosa vs el índice de masa corporal.

Tomando toda esta información podemos indicar que los exámenes realizados tienen anomalías que se deben a los procesos que involucran principalmente la medición de la concentración de la glucosa, ya que esta puede influir en otros aspectos del examen, por lo que es necesario mejorar en dicho aspecto o buscar el motivo por el cual se presentan estas anomalías. Es necesario tomar medidas para corregir este comportamiento, para asegurar la confiabilidad de dicho examen, ya sea a nivel del profesional que realiza el examen o bien sea en el proceso como tal.

EJEMPLO PRÁCTICO CON PENTAHO

A continuación, procederemos a ver una herramienta distinta de minería de datos que se llama Pentaho, esta herramienta nos permitirá realizar un tipo de minería de datos un poco más enfocada a lo que son los reportes y análisis de datos de negocio. Esta herramienta permite a los usuarios más inexpertos realizar un análisis de los datos existentes en una base de datos con el fin de obtener información relevante de esta.

4.5. Paso 1. Identificar las necesidades

Para este ejemplo buscamos desarrollar reportes con la información de ventas detallada de diferentes clientes, con el objetivo de descubrir las tendencias de compra de los clientes, así como los productos con mayor aceptación, posteriormente también se buscará conocer los niveles de venta de cada zona en USA donde se encuentra la compañía, con el

objetivo de mejorar las estrategias de venta y distribución de productos, conocer los puntos fuertes y débiles de ventas.

4.6. Paso 2. Preparación de ambiente y selección de herramientas

Para apreciar de una mejor manera el funcionamiento de Pentaho, utilizaremos una base de datos referente a valores de ventas de una empresa. Que se muestra en la Figura 4.6.1.

Primeramente, para utilizar nuestra herramienta deberemos poseer una base de datos que se encuentre en el formato .csv (valores separados por comas), que es el formato con el cual trabaja Pentaho, como su nombre lo indica simplemente es un archivo que contiene todos los registros de la base de datos con la cual queramos trabajar, separadas por comas.

```
ORDERNUMBER,QUANTITYORDERED,PRICEEACH,ORDERLINENUMBER,SALES,ORDERDATE,STATUS,QTR_ID,MONTH_ID,YEAR_ID,PRODUCTLINE,MSRP,PRODUCTCODE,CUSTOMERNAME,PHONE,ADDRESSLINE1,ADDRESSLINE2,CITY,STATE,POST
10107,30,957,2,2871,2/24/2003 0:00,Shipped,1,2,2003,Motorcycles,95,S10_1678,Land of Toys Inc,2125557818,897 Long Airport Avenue,,NYC,NY,10022,USA,NA,Yu,Kwai,Small
10121,34,8135,5,27659,5/7/2003 0:00,Shipped,2,5,2003,Motorcycles,95,S10_1678,Reims Collectables,26471555,59 rue de l'Abbaye,,Reims,,51100,France,EMEA,Henriot,Paul,Small
10134,41,9474,2,388434,7/1/2003 0:00,Shipped,3,7,2003,Motorcycles,95,S10_1678,Lyon Souveniers,+33 1 46 62 7555,27 rue du Colonel Pierre Avia,,Paris,,75508,France,EMEA,Da Cunha,Daniel,Medium
10145,45,8326,6,37467,8/25/2003 0:00,Shipped,3,8,2003,Motorcycles,95,S10_1678,Toys4GrownUpscom,6265557265,78934 Hillside Dr.,,Pasadena,CA,90003,USA,NA,Young,Julie,Medium
10159,49,100,14,520527,10/10/2003 0:00,Shipped,4,10,2003,Motorcycles,95,S10_1678,Corporate Gift Ideas Co,6505551386,7734 Strong St.,San Francisco,CA,,USA,NA,Brown,Julie,Medium
10168,36,9666,1,347976,10/28/2003 0:00,Shipped,4,10,2003,Motorcycles,95,S10_1678,Technics Stores Inc,6505556809,9408 Furth Circle,,Burlingame,CA,94217,USA,NA,Hirano,Juri,Medium
10180,29,8613,9,249777,11/11/2003 0:00,Shipped,4,11,2003,Motorcycles,95,S10_1678,Daedalus Designs Imports,20161555,"184, chausse de Tournai",,Lille,,59000,France,EMEA,Rance,Martine,Small
10188,48,100,1,551232,11/18/2003 0:00,Shipped,4,11,2003,Motorcycles,95,S10_1678,Herkku Gifts,+47 2267 3215,"Drammen 121, PR 744 Sentrum",,Bergen,,N 5804,Norway,EMEA,Oeztan,Veysel,Medium
10201,22,9857,2,216854,12/1/2003 0:00,Shipped,4,12,2003,Motorcycles,95,S10_1678,Mini Wheels Co,6505555787,5557 North Pendale Street,San Francisco,CA,,USA,NA,Murphy,Julie,Small
10211,41,100,14,470844,1/15/2004 0:00,Shipped,1,1,2004,Motorcycles,95,S10_1678,Auto Canal Petit,(1) 47556555,"25, rue Lauriston",,Paris,,75016,France,EMEA,Perrier,Dominique,Medium
10223,37,100,1,396566,2/20/2004 0:00,Shipped,1,2,2004,Motorcycles,95,S10_1678,"Australian Collectors, Co",03 9520 4555,636 St Kilda Road,Level 3,Melbourne,Victoria,3004,Australia,APAC,Ferguson,Peter,Medium
10237,23,100,7,233312,4/5/2004 0:00,Shipped,2,4,2004,Motorcycles,95,S10_1678,Vitachrome Inc,2125551500,2678 Kingston Rd,Suite 101,NYC,NY,10022,USA,NA,Frick,Michael,Small
10251,28,100,2,318864,5/18/2004 0:00,Shipped,2,5,2004,Motorcycles,95,S10_1678,Tekni Collectables Inc,2015559350,7476 Moss Rd.,Newark,NJ,94019,USA,NA,Brown,William,Medium
10263,34,100,2,367676,6/28/2004 0:00,Shipped,2,6,2004,Motorcycles,95,S10_1678,Gift Depot Inc,2035552570,25593 South Bay Ln,,Bridgewater,CT,97562,USA,NA,King,Julie,Medium
10275,45,9283,1,417735,7/23/2004 0:00,Shipped,3,7,2004,Motorcycles,95,S10_1678,La Rochelle Gifts,40678555,"67, rue des Cinquante Otages",,Nantes,,44000,France,EMEA,Labrune,Janine,Medium
10285,36,100,6,409968,8/27/2004 0:00,Shipped,3,8,2004,Motorcycles,95,S10_1678,Marta's Replicas Co,6175558555,39323 Spinnaker Dr,,Cambridge,MA,51247,USA,NA,Hernandez,Marta,Medium
10299,23,100,9,259739,9/30/2004 0:00,Shipped,3,9,2004,Motorcycles,95,S10_1678,"Toys of Finland, Co",90-224 8555,Keskuskatu 45,,Helsinki,,21240,Finland,EMEA,Karttunen,Matti,Small
10309,41,100,5,439438,10/15/2004 0:00,Shipped,4,10,2004,Motorcycles,95,S10_1678,Baane Mini Imports,07-98 9555,Erling Skakkes gate 78,,Stavern,,4110,Norway,EMEA,Bergulfsen,Jonas,Medium
10318,46,9474,1,435804,11/2/2004 0:00,Shipped,4,11,2004,Motorcycles,95,S10_1678,Diecast Classics Inc,2155551555,7586 Pompton St.,Allentown,PA,70267,USA,NA,Yu,Kyung,Medium
10329,42,100,1,439614,11/15/2004 0:00,Shipped,4,11,2004,Motorcycles,95,S10_1678,Land of Toys Inc,2125557818,897 Long Airport Avenue,,NYC,NY,10022,USA,NA,Yu,Kwai,Medium
10341,41,100,5,773793,11/24/2004 0:00,Shipped,4,11,2004,Motorcycles,95,S10_1678,Salzburg Collectables,6562-9555,Geislweg 14,,Salzburg,,5020,Austria,EMEA,Pipps,Georg,Large
10361,20,7255,13,1451,12/17/2004 0:00,Shipped,4,12,2004,Motorcycles,95,S10_1678,Souveniers And Things Co,+61 2 9495 8555,"Monitor Money Building, 815 Pacific Hwy",,Level 6,Chatswood,NSW,2067,Australia,APAC,Huxley,Adrian,Small
10375,21,3491,12,73311,2/3/2005 0:00,Shipped,1,2,2005,Motorcycles,95,S10_1678,La Rochelle Gifts,40678555,"67, rue des Cinquante Otages",,Nantes,,44000,France,EMEA,Labrune,Janine,Small
10388,42,7636,4,320712,3/3/2005 0:00,Shipped,1,3,2005,Motorcycles,95,S10_1678,FunGiftIdeascom,5085552555,1785 First Street,,New Bedford,MA,50553,USA,NA,Benitez,Violeta,Medium
10403,24,100,7,243456,4/8/2005 0:00,Shipped,2,4,2005,Motorcycles,95,S10_1678,"UK Collectables, Ltd",(171) 555-2282,Berkeley Gardens 12  Brewery,,Liverpool,,WX1 6LT,UK,EMEA,Devon,Elizabeth,Small
10417,66,100,2,751608,5/13/2005 0:00,Disputed,2,5,2005,Motorcycles,95,S10_1678,Euro Shopping Channel,(91) 555 94 44,"C/ Moralzarzal, 86",,Madrid,,28034,Spain,EMEA,Freyre,Diego,Large
10103,26,100,11,540462,1/29/2003 0:00,Shipped,1,1,2003,Classic Cars,214,S10_1949,Baane Mini Imports,07-98 9555,Erling Skakkes gate 78,,Stavern,,4110,Norway,EMEA,Bergulfsen,Jonas,Medium
10112,29,100,1,720911,3/24/2003 0:00,Shipped,1,3,2003,Classic Cars,214,S10_1949,"Volvo Model Replicas, Co",0921-12 3555,Berguvsv,gen  8,,Lule,,S-958 22,Sweden,EMEA,Berglund,Christina,Large
10126,38,100,11,732906,5/28/2003 0:00,Shipped,2,5,2003,Classic Cars,214,S10_1949,"Corrida Auto Replicas, Ltd",(91) 555 22 82,"C/ Araquil, 67",,Madrid,,28023,Spain,EMEA,Sommer,Martin,Large
10140,37,100,11,73741,7/24/2003 0:00,Shipped,3,7,2003,Classic Cars,214,S10_1949,Technics Stores Inc,6505556809,9408 Furth Circle,,Burlingame,CA,94217,USA,NA,Hirano,Juri,Large
10150,45,100,8,109935,9/19/2003 0:00,Shipped,3,9,2003,Classic Cars,214,S10_1949,"Dragon Souveniers, Ltd",+65 221 7555,"Bronz Sok, Bronz Apt 3/6 Tesvikiye",,Singapore,,79903,Singapore,Japan,Natividad,Eric,Large
10163,21,100,1,486024,10/20/2003 0:00,Shipped,4,10,2003,Classic Cars,214,S10_1949,Classic Legends Inc,2125558493,5905 Pompton St,Suite 750,NYC,NY,10022,USA,NA,Hernandez,Maria,Medium
10174,34,100,4,801482,11/6/2003 0:00,Shipped,4,11,2003,Classic Cars,214,S10_1949,"Australian Gift Network, Co",61-7-3844-6555,31 Duncan St West End,,South Brisbane,Queensland,4101,Australia,APAC,Calaghan,Tony,Large
10183,23,100,8,537257,11/13/2003 0:00,Shipped,4,11,2003,Classic Cars,214,S10_1949,"Classic Gift Ideas, Inc",2155554695,782 First Street,,Philadelphia,PA,71270,USA,NA,Cervantes,Francisca,Medium
10194,42,100,11,729036,11/25/2003 0:00,Shipped,4,11,2003,Classic Cars,214,S10_1949,"Saveley & Henriot, Co",78325555,"2, rue du Commerce",,Lyon,,69004,France,EMEA,Saveley,Mary,Large
10206,47,100,6,906489,12/5/2003 0:00,Shipped,4,12,2003,Classic Cars,214,S10_1949,Canadian Gift Exchange Network,(604) 555-3392,1900 Oak St.,Vancouver,BC,V3F 2K1,Canada,NA,Tannamuri,Yoshi,Large
10215,35,100,3,60753,1/29/2004 0:00,Shipped,1,1,2004,Classic Cars,214,S10_1949,West Coast Collectables Co,3105553722,3675 Furth Circle,,Burbank,CA,94019,USA,NA,Thompson,Steve,Medium
10228,29,100,2,646323,3/10/2004 0:00,Shipped,1,3,2004,Classic Cars,214,S10_1949,Cambridge Collectables Co,6175555555,4658 Baden Av,,Cambridge,MA,51247,USA,NA,Tseng,Kyung,Medium
```

Figura 4.6.1. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

Con este archivo procederemos a cargarlo en nuestra plataforma de Pentaho presentada en la Figura 4.6.2., para este ingresaremos a nuestra herramienta la cual nos presentará la siguiente pantalla.

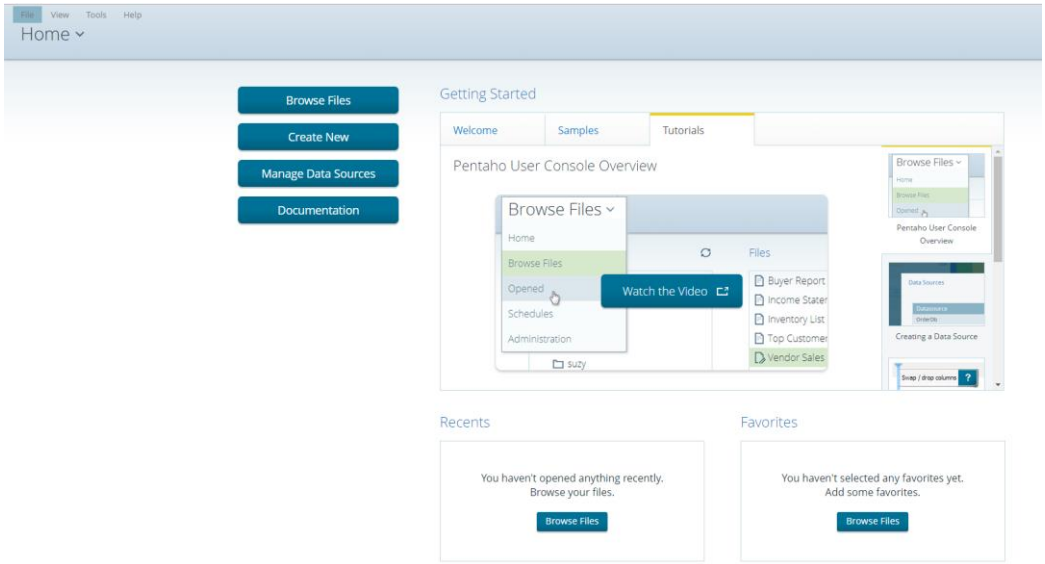


Figura 4.6.2. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

Esta pantalla será la que nos permita ingresar diferentes bases de datos, crear reportes, reportes interactivos y dashboard. Para cargar los registros de nuestro archivo csv procedemos a seleccionar la opción “Manage Data Source”, la cual nos presentara la siguiente pantalla, como se ve en la Figura 4.6.3.

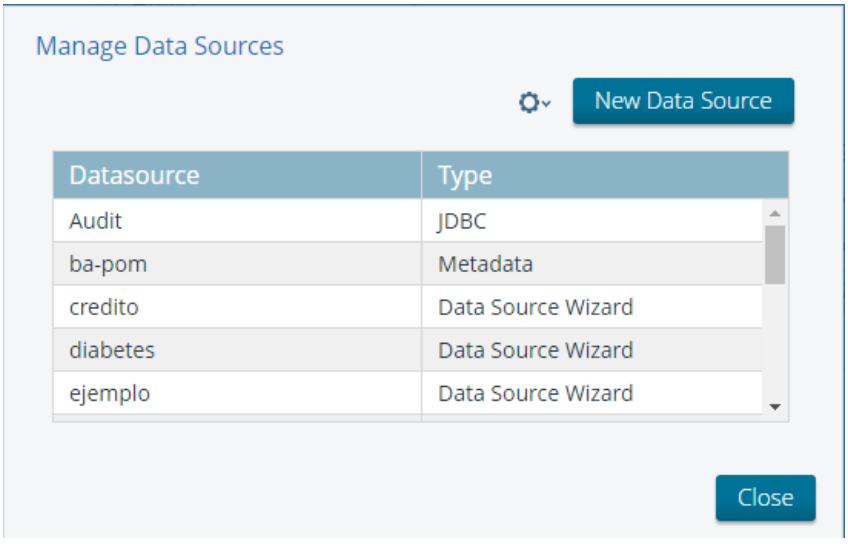


Figura 4.6.3. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

En este recuadro podemos apreciar las diferentes fuentes de datos disponibles para utilizarse con nuestra herramienta, para nuestro ejemplo

crearemos una nueva fuente de datos. Seleccionamos la opción “New Data Source”, como se presenta a continuación en la Figura 4.6.4.

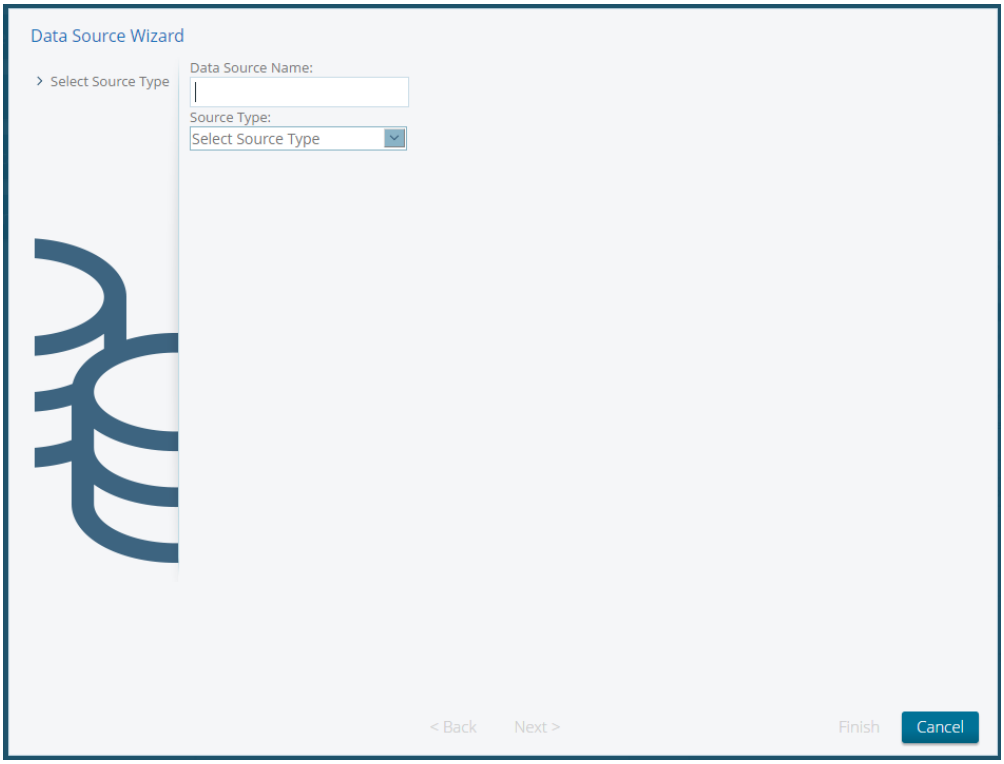


Figura 4.6.4. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

En esta pantalla seleccionaremos nuestro tipo de archivo, en nuestro caso CSV, y pondremos un nombre a esta nueva fuente de datos, en nuestro caso colocaremos en nombre de “Ventas”. Una vez ingresado estos datos se habilitan más opciones que nos permitirán buscar el origen de nuestros datos, así como una vista previa de los datos. Como se ve en la Figura 4.6.5.

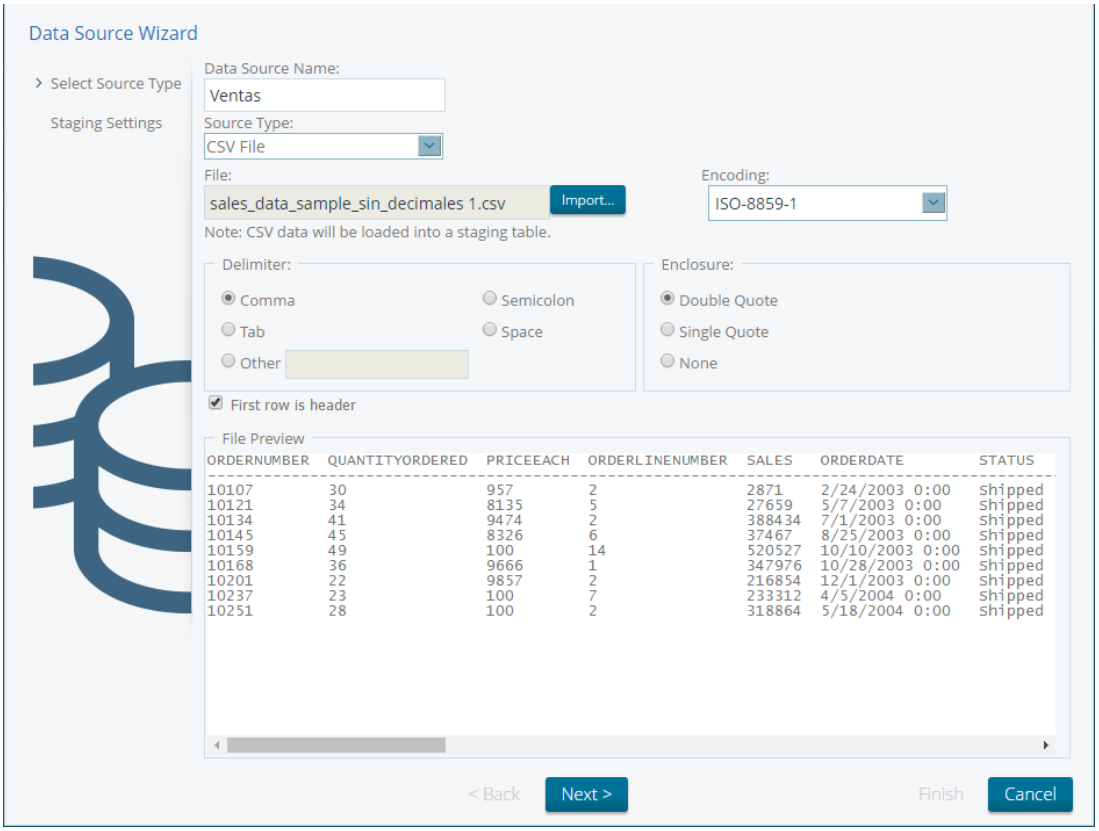


Figura 4.6.5. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

Procedemos a la siguiente pantalla, donde podremos personalizar, si así lo deseamos, los tipos de datos que se reconocen al momento de cargar nuestros datos. Como todo parece estar correcto dejaremos la configuración por defecto. Como se observa en la Figura 4.6.6.

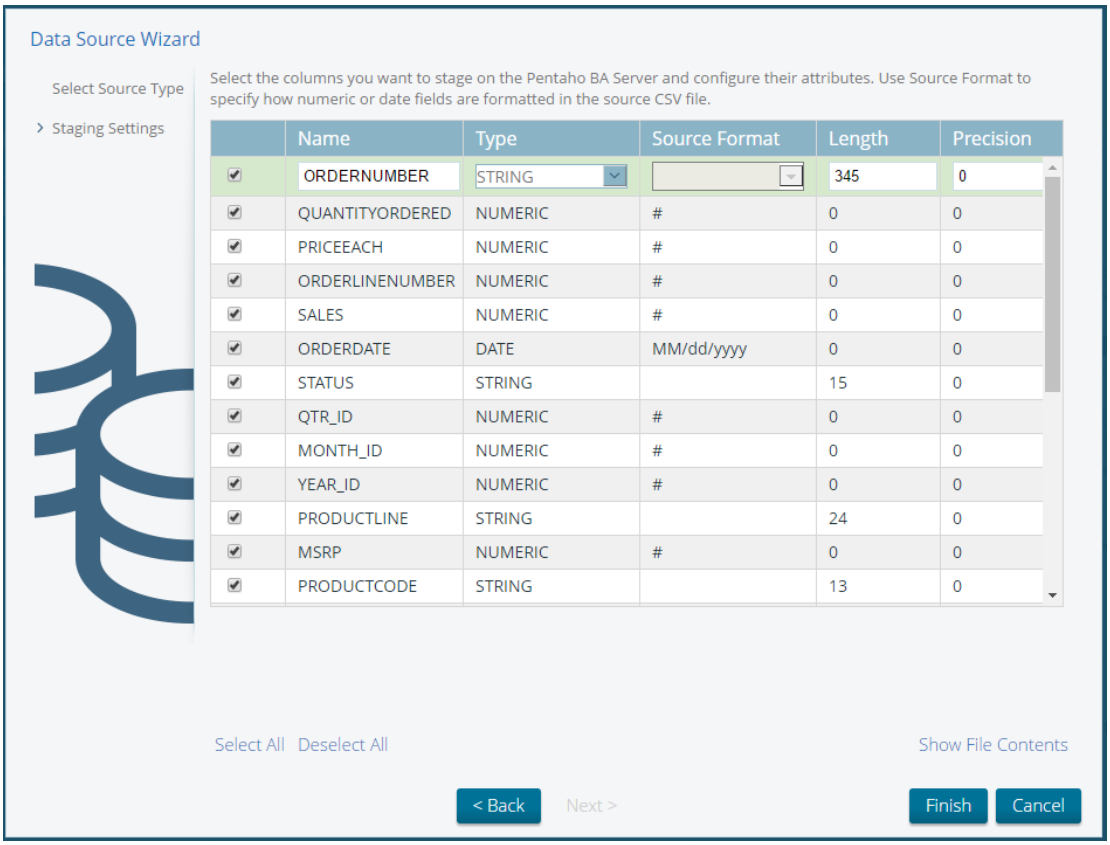


Figura 4.6.6. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

Procedemos a finalizar nuestra carga de datos, presionando la opción “Finish”, si no han existido problemas se presentará un cuadro informándonos que todos los registros han sido ingresados, como se ve en la Figura 4.6.7.

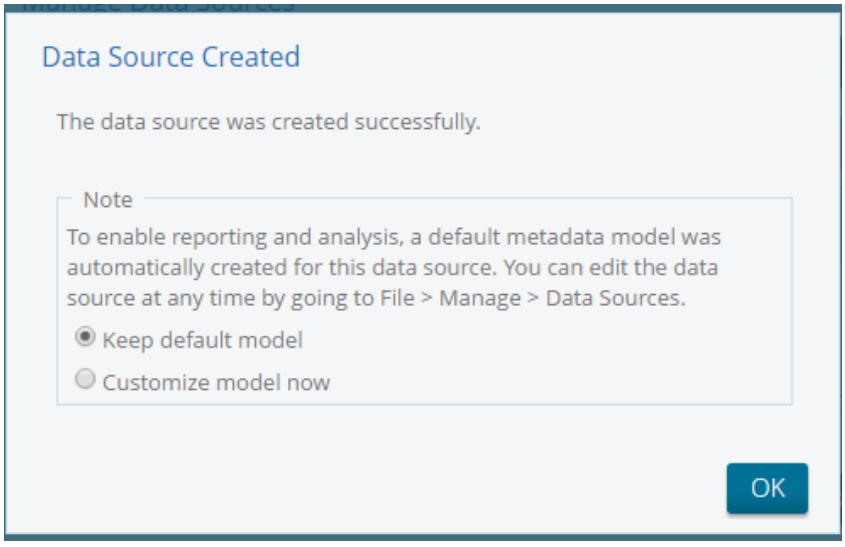


Figura 4.6.7. Paso 2, Elaborado por: Oscar Córdova y Carlos Rosales

Con esto hemos logrado exitosamente subir todos nuestros registros, a continuación, empezaremos con la creación de reportes. Esta opción es una de las más utilizadas por los trabajadores de una organización, debido principalmente a que los reportes son el medio principal para defender muchos de los proyectos de la organización, este tipo de reportes nos permite realizar un análisis de los datos, permitiendo conocer así el estado de un proyecto.

4.7. Paso 3. Empezando con la minería de datos.

4.7.1. Reportes Interactivos.

Para iniciar con nuestro reporte, regresaremos a la pantalla principal de nuestra herramienta Pentaho, donde seleccionaremos la opción “Create New”, seguido de la opción “Interactive Report”. Como se observa de la Figura 4.7.1.1.

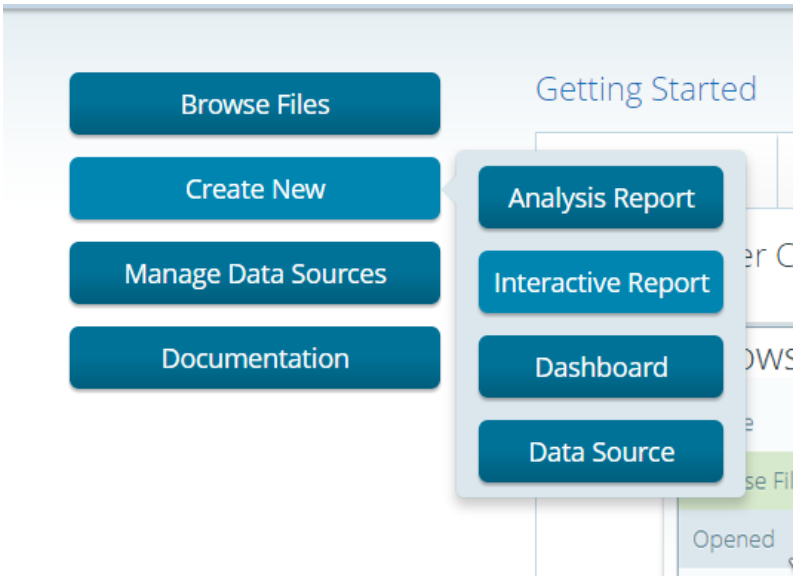


Figura 4.7.1.1. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

A parecerá ante nosotros un recuadro como se observa en la Figura 4.7.1.2, donde podremos seleccionar la fuente de datos que utilizaremos para crear nuestro reporte, en nuestro caso seleccionaremos la fuente de datos que acabamos de crear “Ventas”.

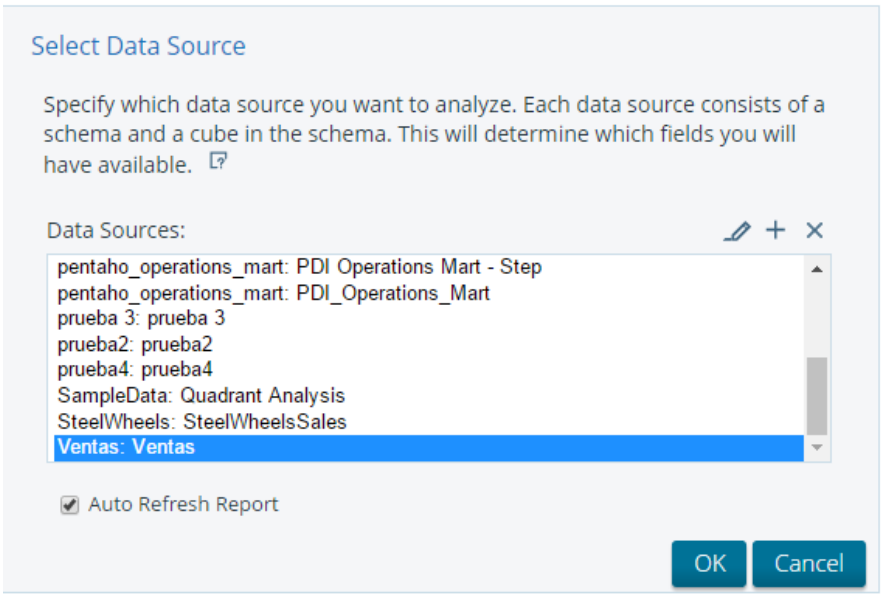


Figura 4.7.1.2. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Una vez seleccionada nuestra fuente de datos se nos presentara la pantalla donde procederemos a crear nuestro reporte. Como se ve en la Figura 4.7.1.3.

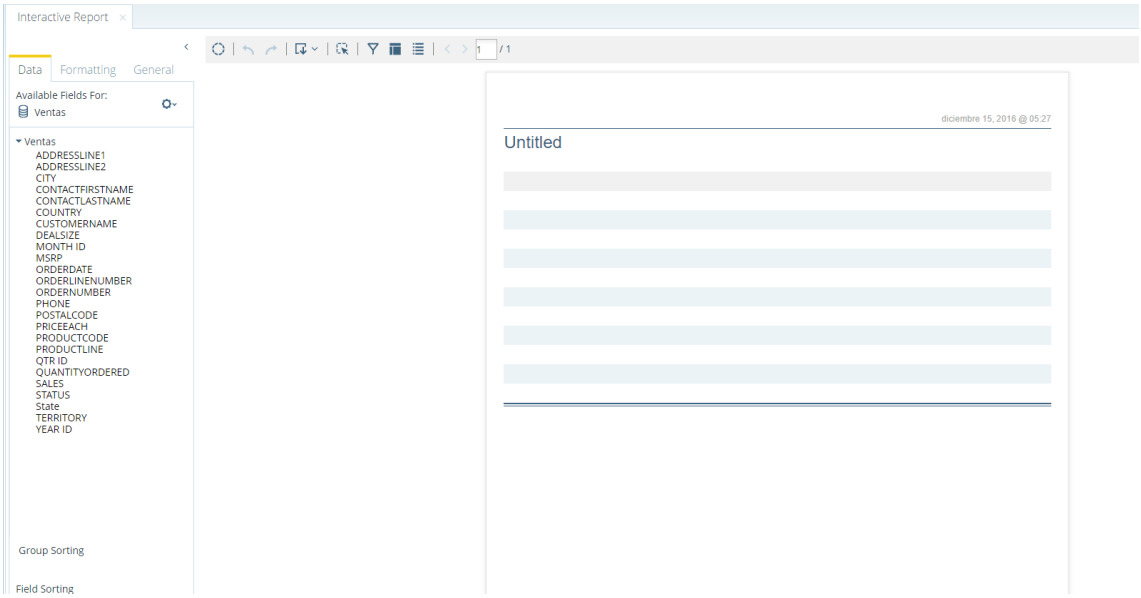


Figura 4.7.1.3. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Podemos apreciar que todos los campos de nuestra base de datos se encuentran en la parte izquierda de la pantalla, donde podremos arrastrar dichos campos a la parte central donde existe una plantilla para nuestro

reporte. Al momento de crear nuestro reporte debemos tomar en consideración el objetivo que tratamos de cumplir. Esto nos permitirá tener una mayor claridad de los campos y tipos de datos que deberemos utilizar en nuestro reporte, con el fin de expresar de mejor manera nuestro objetivo.

Para nuestro caso buscaremos realizar un informe de artículos y sus precios clasificado por los nombres de los compradores. Este reporte lo nombraremos reporte de compra.

Arrastrando el atributo de “Product Line” de manera horizontal hacia nuestra plantilla de reporte, posteriormente podemos seleccionar los atributos que deseemos que aparezcan de cada producto como su código de producto, el precio y la fecha de compra. Finalmente podemos aplicar el filtro por cliente, realizando clic derecho sobre el atributo “CustomerName” y seleccionando “Prompt”, nuestro reporte debería lucir así. El reporte se lo puede apreciar en la Figura 4.7.1.4.

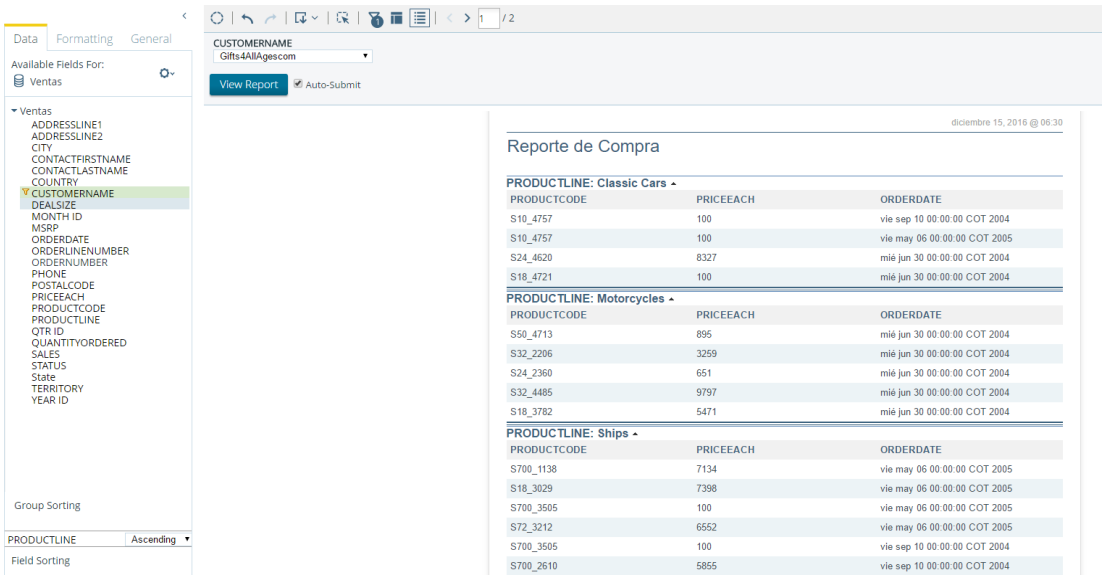


Figura 4.7.1.4. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Hemos logrado crear un reporte de compra de cada uno de los clientes, independiente de cada uno de ellos. En este caso podemos apreciar que se han comprado varios juguetes como carros clásicos, motocicletas, barcos y trenes, con sus respectivos precios, por parte del cliente “Gifts4AllAges”.

Si cambiamos el filtro de nuestro cliente el reporte también cambiará. Como se presenta a continuación en la Figura 4.7.1.5.

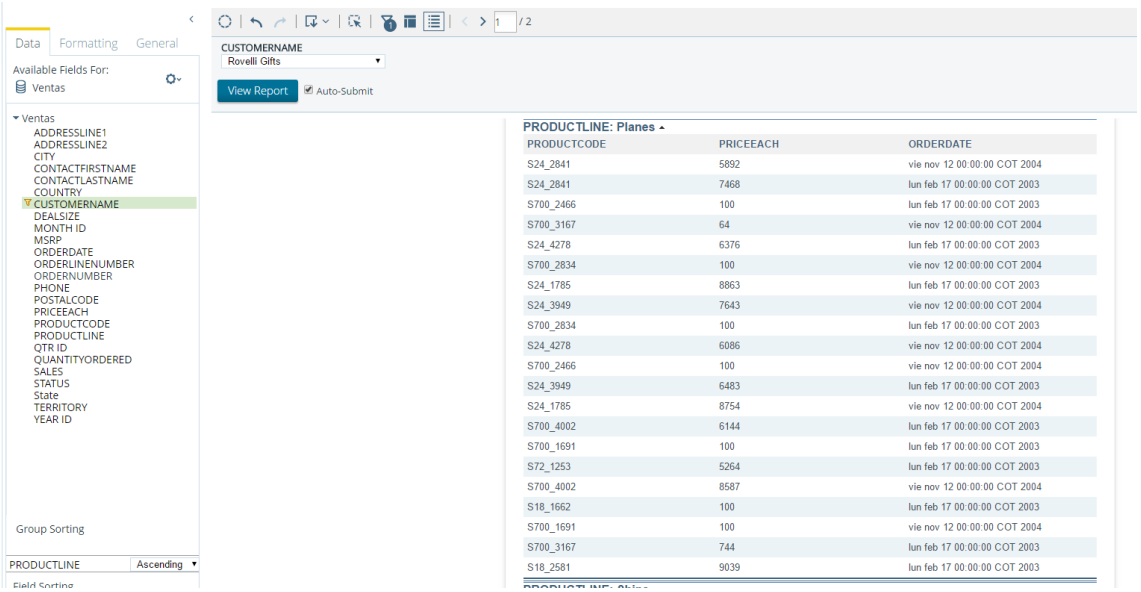


Figura 4.7.1.5. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Este tipo de reporte nos da como evidencia todas las compras realizadas por un cliente a lo largo del tiempo, en este caso se aprecia que el cliente “Rovelli Gifts” ha comprado una gran cantidad de aviones de juguete, si añadimos una columna más a nuestro reporte podremos apreciar la cantidad de aviones comprada en cada modelo. Como se ve en la Figura 4.7.1.6.

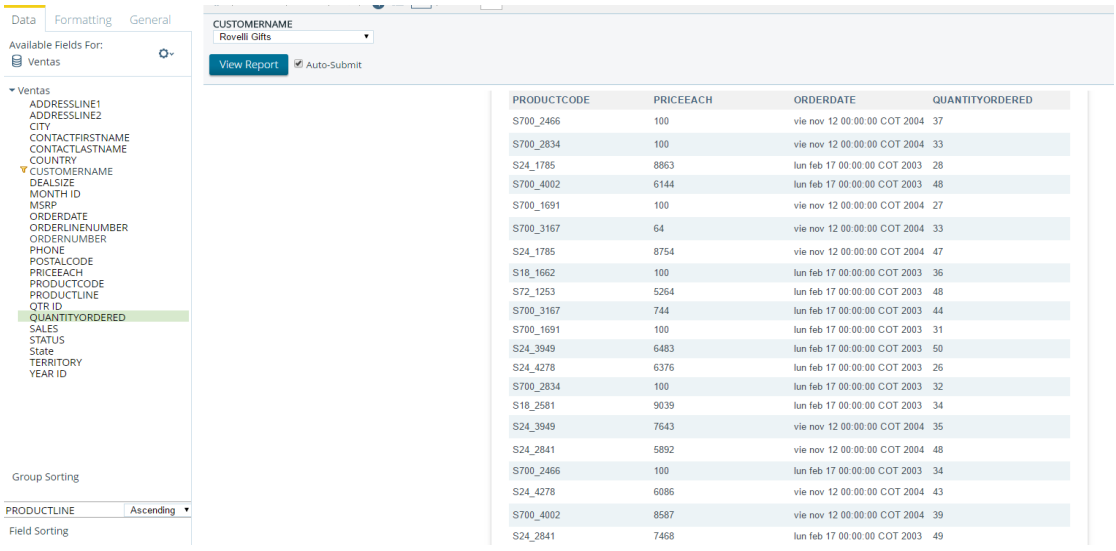


Figura 4.7.1.6. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

4.7.2. Reportes de análisis.

Otra de las funcionalidades principales de Pentaho que es útil para realizar una especie de minería de datos son los “Analysis Reports”, reportes de datos más congruentes que nos permitirán realizar cruces de la información, con el fin de obtener información específica.

En nuestra pantalla principal de Pentaho, una vez más, seleccionaremos la opción “Create New” posteriormente “Analysis Report”. Como se ve en la Figura 4.7.2.1.

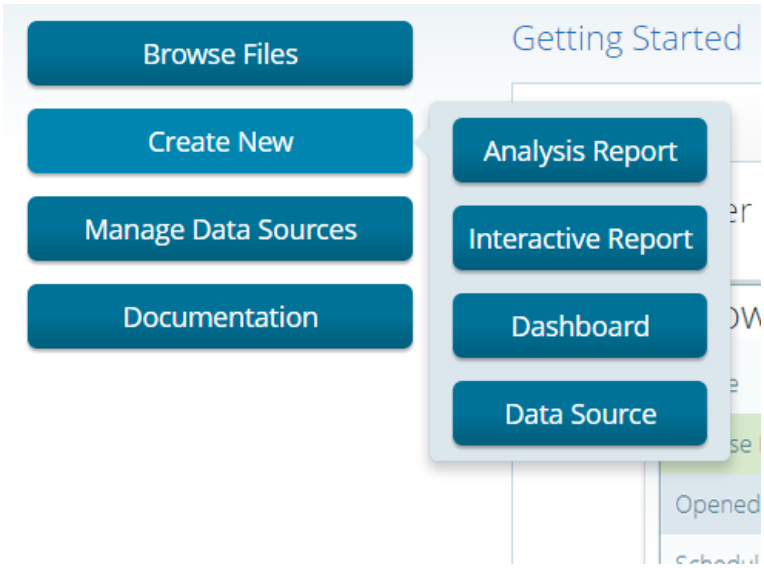


Figura 4.7.2.1. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Una vez más seleccionaremos como fuente de datos la opción “Ventas”, que creamos anteriormente. Como se aprecia en la Figura 4.7.2.2.

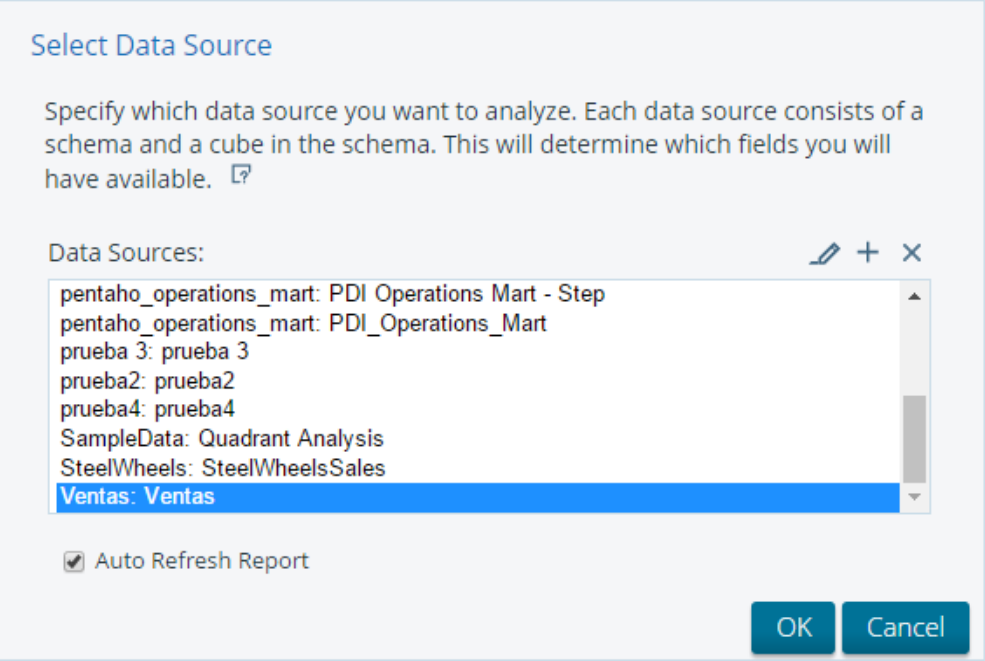


Figura 4.7.2.2. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

En la pantalla a continuación que se presenta en la Figura 4.7.2.3. podemos observar nuestros atributos, la sección donde se presentarán los datos, y los campos que nos permitirán realizar los cruces de atributos. En este caso buscaremos cuales son los totales de ventas a lo largo del país “USA”, con el fin de conocer el territorio con mayores ganancias para la compañía, de mismo modo el que menor ganancia genera.

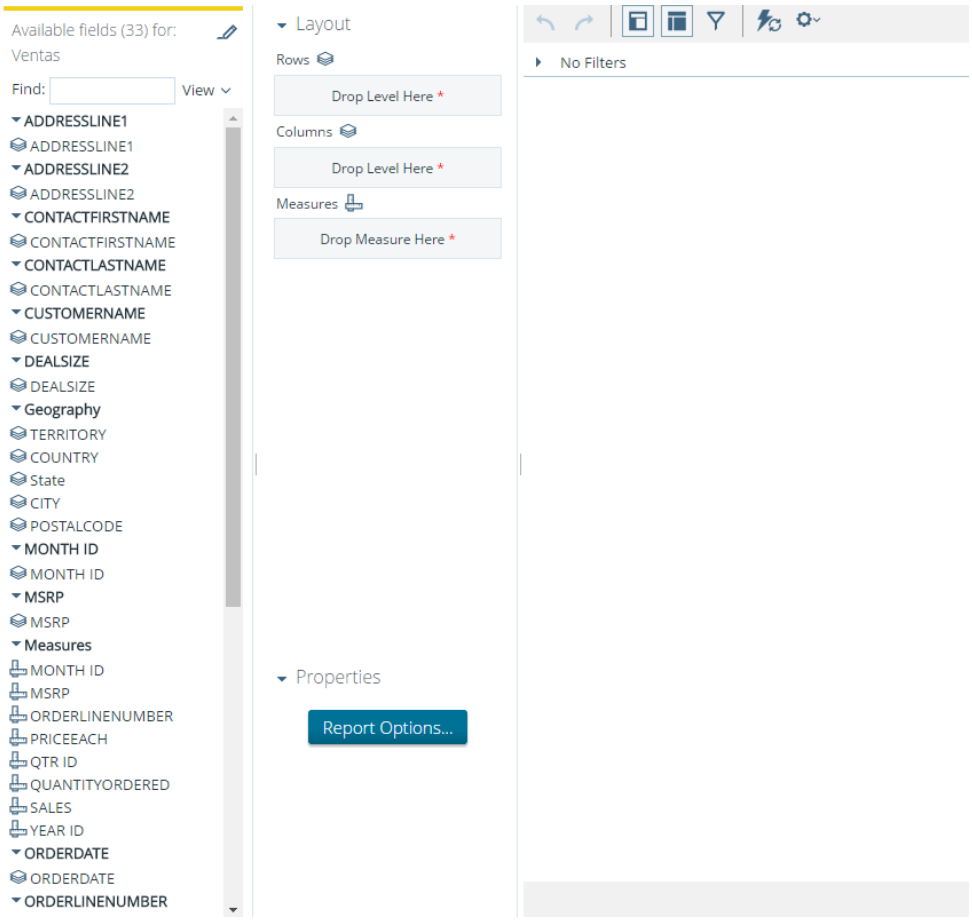


Figura 4.7.2.3. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Ya que nuestra variable más importante en este caso son las ventas totales, procederemos a arrastrar el atributo “Sales”, a nuestro campo “Measures”. Esto permitirá que los cruces de atributos que se realicen tengan como información principal las ventas. Como se observa en la Figura 4.7.2.4.

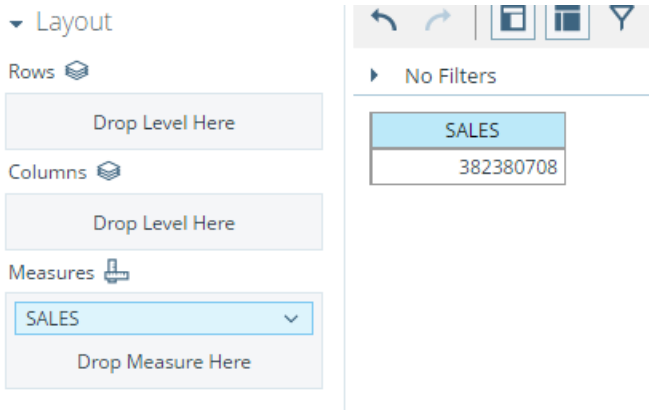


Figura 4.7.2.4. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Podemos apreciar en la Figura 4.7.2.5. que al hacer esto las ventas totales de la compañía aparecen en la parte derecha de la pantalla. Ya que nuestro objetivo es conocer las ventas totales por territorio, procederemos a arrastrar el atributo “Territory” al campo “Rows”, podemos ver que las ventas totales se han distribuido en los diferentes países en los cuales se encuentra la compañía.

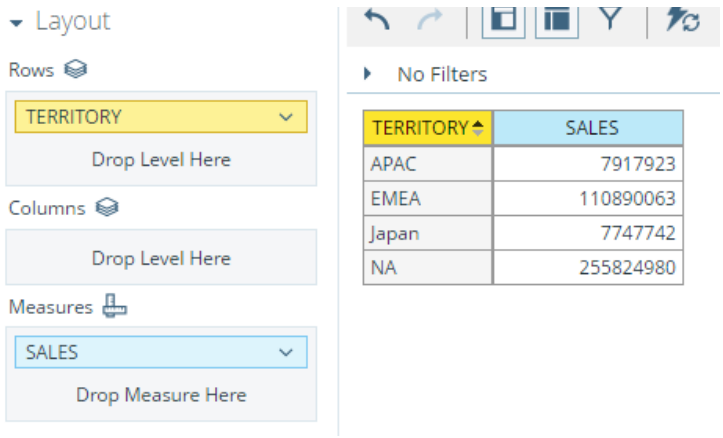


Figura 4.7.2.5. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Como nuestro objetivo es conocer en que territorio específico de “USA” es donde más ventas se realizan es necesario desglosar todavía más nuestra información, para esto simplemente realizaremos doble clic sobre el campo “NA” (Norte América), lo que nos desplegara la información de los países pertenecientes a este territorio. Como se observa en la Figura 4.7.2.6.

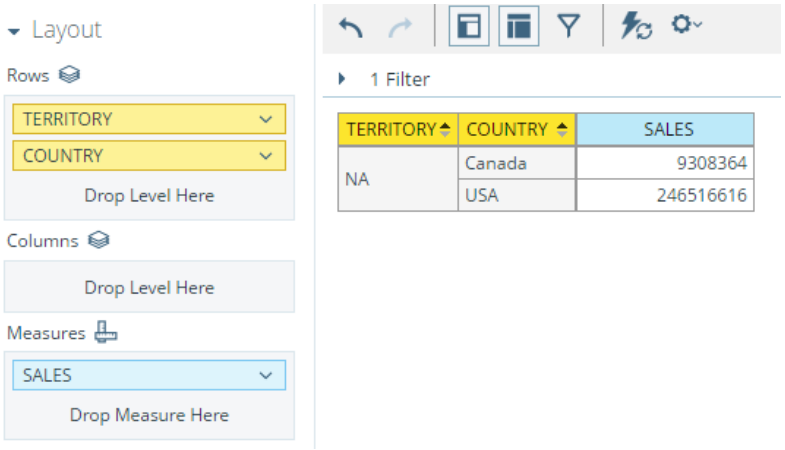
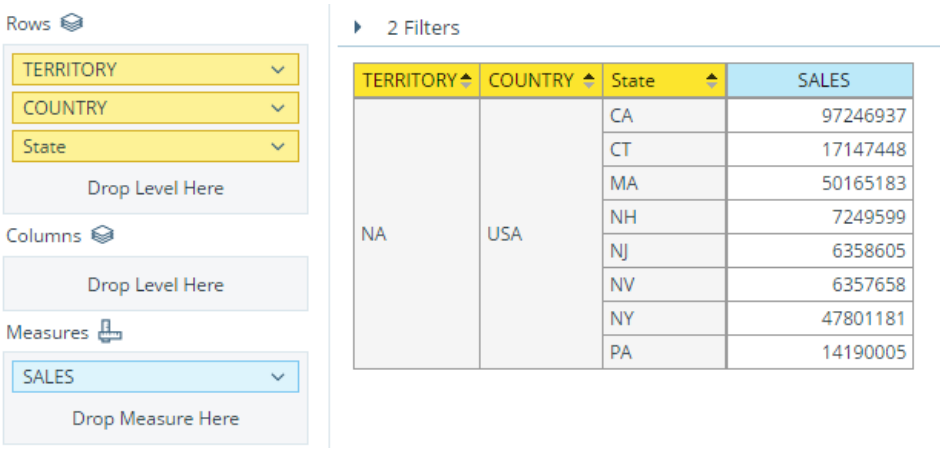


Figura 4.7.2.6. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

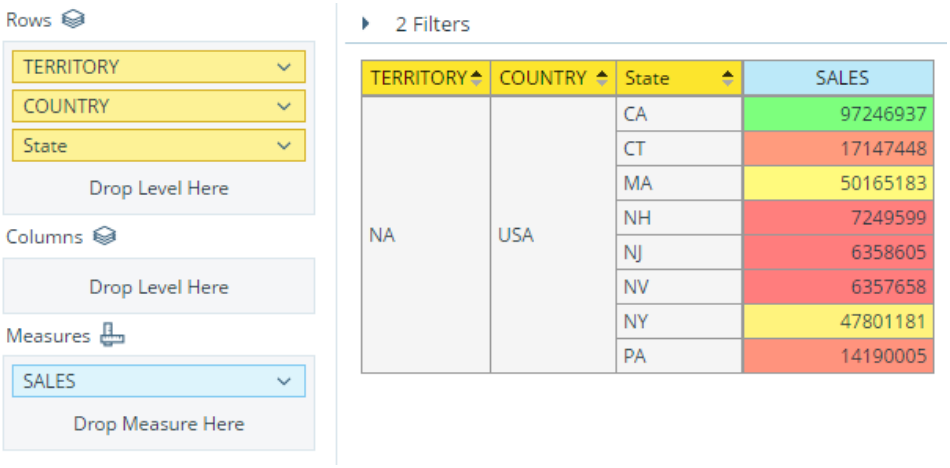
Procedemos a seleccionar “USA” para que se desplieguen los estados donde se encuentra la compañía. Como se ve en la Figura 4.7.2.7.



TERRITORY	COUNTRY	State	SALES
NA	USA	CA	97246937
		CT	17147448
		MA	50165183
		NH	7249599
		NJ	6358605
		NV	6357658
		NY	47801181
		PA	14190005

Figura 4.7.2.7. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Pentaho nos ofrece una opción de paleta de colores que nos permitirá distinguir con facilidad el valor más alto y el más bajo. Como se observa en la Figura 4.7.2.8.



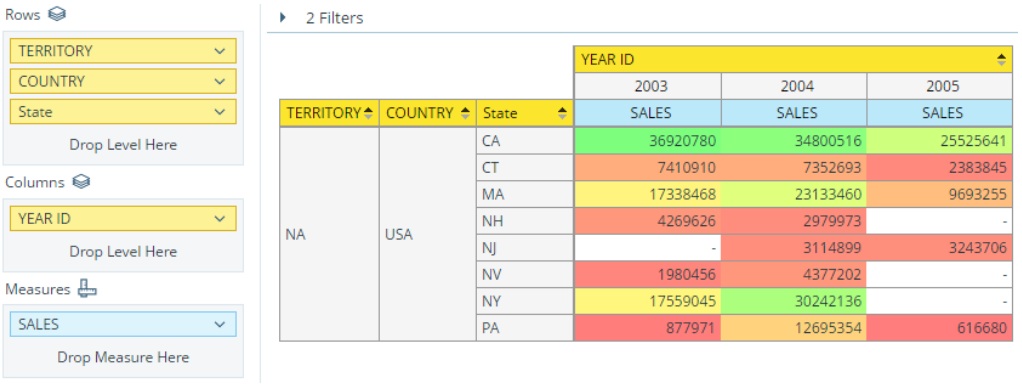
TERRITORY	COUNTRY	State	SALES
NA	USA	CA	97246937
		CT	17147448
		MA	50165183
		NH	7249599
		NJ	6358605
		NV	6357658
		NY	47801181
		PA	14190005

Figura 4.7.2.8. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

Podemos observar que el estado con mayores ventas totales es el estado de California, mientras que el estado con ventas más bajas es Nevada. Si nos basamos en estos datos podemos detectar que existen problemas de venta no simplemente en el estado de Nevada, el estado de New Hamsphire y New Jersey se encuentran con niveles muy bajos de

ganancias, por lo que sería necesario realizar una nueva estrategia de ventas para estos estados, con el fin de elevar las ganancias.

Para realizar un análisis más detallado sobre las ganancias, podemos colocar un filtro más que nos permita ver las ganancias a través de los años. Como se observa en la Figura 4.7.2.9.



			YEAR ID		
			2003	2004	2005
TERRITORY	COUNTRY	State	SALES	SALES	SALES
NA	USA	CA	36920780	34800516	25525641
		CT	7410910	7352693	2383845
		MA	17338468	23133460	9693255
		NH	4269626	2979973	-
		NJ	-	3114899	3243706
		NV	1980456	4377202	-
		NY	17559045	30242136	-
		PA	877971	12695354	616680

Figura 4.7.2.9. Paso 3, Elaborado por: Oscar Córdova y Carlos Rosales

4.8. Paso 4. Análisis e interpretación de resultados.

En este paso se interpreta los resultados obtenidos con la herramienta de Pentaho en los literales 4.7.1. Reporte Interactivos y 4.7.2. Reporte de Análisis.

4.8.1. Reportes Interactivos.

Otro fin de este tipo de reportes es el de investigación de las tendencias de compra de los clientes de la organización, permitiéndonos desarrollar estrategias de ventas personalizadas para nuestros clientes, esto en base a sus registros de compra. Por ejemplo, al cliente “Rovelli Gifts” se les ofrecerían promociones referentes a los juguetes de avión, en cambio al cliente “Gifts4AllAges” sería oportuno enviarle varios tipos de promociones, esto debido a que sus tendencias de compra son más variadas.

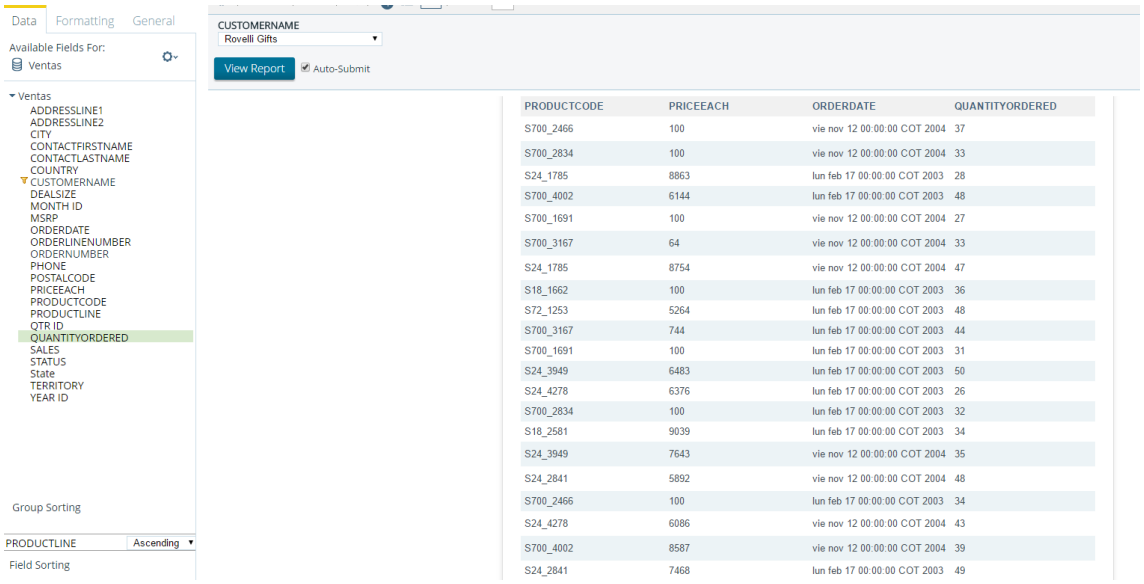
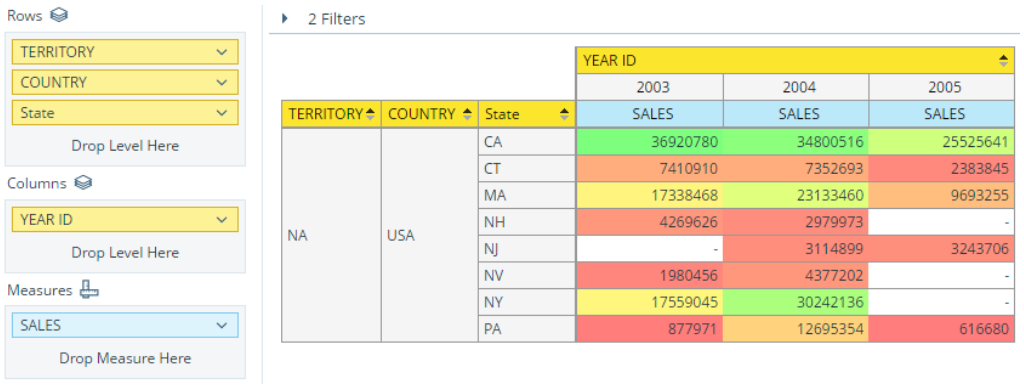


Figura 4.8.1.1. Paso 4 Elaborado por: Oscar Córdova y Carlos Rosales

4.8.2. Reportes de análisis.

Con este nuevo cuadro podemos observar que a pesar de que California se mantiene como el estado con mayores ventas a través de los años, vemos que las ganancias empiezan a bajar año tras año, lo que podría representar un problema para la compañía a futuro, por lo que es necesario investigar los motivos de esta baja en las ganancias y corregirlo.



TERRITORY	COUNTRY	State	YEAR ID 2003 SALES	2004 SALES	2005 SALES
NA	USA	CA	36920780	34800516	25525641
		CT	7410910	7352693	2383845
		MA	17338468	23133460	9693255
		NH	4269626	2979973	-
		NJ	-	3114899	3243706
		NV	1980456	4377202	-
		NY	17559045	30242136	-
		PA	877971	12695354	616680

Figura 4.8.2.1. Paso 4, Elaborado por: Oscar Córdova y Carlos Rosales

5. CAPÍTULO 5: CONCLUSIONES Y RECOMENDACIONES

En el siguiente capítulo se presentará las conclusiones y recomendaciones que hemos encontrado en el desarrollo del trabajo de disertación.

5.1. CONCLUCIONES

- Los resultados del proceso de minería de datos pueden variar dependiendo de la herramienta utilizada en el proceso, la explicación de los resultados está sujeta a la interpretación del encargado de la minería; por lo tanto, es aconsejable el previo conocimiento teórico del proceso de minado de datos con el fin de detectar los posibles errores de interpretación que se presenten en los resultados, como para la mejor justificación de los mismos resultados.
- La preparación de la base de datos, como de la superficie minable son partes cruciales del proceso de minería; la selección de la información relevante debe hacerse con precaución siempre teniendo en cuenta el objetivo que se desea cumplir, mejorando de esta manera los resultados obtenidos.
- El proceso de minería está sujeto a muchos tipos de errores, ya sea la incompatibilidad de los tipos de datos con el algoritmo de minería elegido o resultados incoherentes con respecto a la información utilizada; por lo que es necesario cumplir con todos los pasos previstos para realizar una buena minería de datos y minimizar la cantidad de errores que se presentaran en los resultados finales.
- Para la realización de minería de datos es esencial el conocimiento previo de varios conceptos básicos, para poder familiarizarse con el tema, por lo que es importante el aprender o entender sobre los métodos que existen para el minado de datos en relación con grandes cantidades de información almacenada en las bases de datos de cualquier tipo de organización.
- La minería de datos nos ayuda mucho a la toma decisiones en cualquier campo laboral, ya que toma una gran cantidad de

información la evalúa o la compara según las necesidades del usuario final generando así la mejor ruta o la respuesta más óptima para el problema presentado.

- La información que es generada en el minado de datos puede no solo constar de una sola solución, ya que puede generar varios tipos de soluciones a los problemas presentados por las organización o usuario final, adicionalmente en caso de no encontrar una solución basándose en los datos históricos y actuales puede crear o generar una nueva solución al problema presentado.
- Pentaho es una herramienta que permite a los usuarios con conocimiento básico el analizar datos, en grandes cantidades, obteniendo información relevante que permitirá tomar decisiones o reconocer tendencias.
- A pesar de que es una herramienta más simple que WEKA, Pentaho cumple óptimamente con su objetivo al momento de realizar minería de datos, lo que nos demuestra que utilizando los métodos y herramientas adecuadas la minería de datos puede aplicarse sin muchos problemas a la mayoría de organizaciones, empresas con grandes cantidades de datos, ayudando a mejorar no solo las ganancias de la empresa, sino a corregir errores y problemas que parecen complicados a simple vista, pero que en realidad tienen una solución al momento de desglosar la información que se maneja. Grandes cantidades de información son muy complicadas de manejar y más complicado aún es obtener información de dichos datos, por lo que la minería de datos a través de herramientas visuales como Pentaho tienen un enorme potencial para el desarrollo de una organización.
- La herramienta WEKA tiene un gran potencial para realizar diferentes tipos de minería de datos, así como la aplicación de diferentes algoritmos de minería de datos dando lugar a una gran cantidad de posibles resultados, como gráficos representando datos, árboles de decisiones, matriz de decisiones, descubrimiento de relaciones entre atributos, etc. Sin embargo, todos estos resultados son datos totalmente matemáticos y libres a la interpretación de quien realiza la minería de datos. Además muchos de estos resultados necesitan de conocimiento previo de

la base de datos, así como la habilidad de interpretar el significado de matrices de confusión.

- En conclusión, WEKA es una herramienta que nos permite analizar los datos y realizar una gran y minuciosa minería de datos, mas no es apta para que un usuario común y sin previo conocimiento y preparación la use, debido a la dificultad de interpretar los resultados.

5.2. RECOMENDACIONES

- Para el proceso de minado de datos se recomienda el conocimiento previo de fundamentos teóricos sobre bases de datos y de minería de datos; por lo que ayudará a la justificación e interpretación de los resultados obtenidos.
- Para la realización de minería de datos es necesario el preparar una superficie minable dependiendo de las necesidades y objetivos que se desee cumplir.
- Se recomienda que existan calidad en los datos, en donde se considerará el duplicamiento de datos y la forma en la que se encuentra redactada la sintaxis de los datos; para así evitar los posibles errores que puedan generar problemas de compatibilidad de los tipos de datos.
- Para la comprensión de la minería de datos se recomienda comenzar con el aprendizaje de los conceptos básicos de la información, el origen de los datos y las técnicas con sus respectivas metodologías del minado de datos.
- La minería de datos puede generar varias soluciones de calidad por lo que se insinúa que se seleccione la mejor solución tomando en cuenta los datos o la información que fue tomada en cuenta para la realización de este proceso.
- Para la generación de resultados por medio del minado de datos a los problemas que se presenten es necesario que se tengan almacenados datos históricos y actuales sobre las posibles soluciones ya presentadas en otros casos para así generar una

nueva salida que sea de ayuda para la organización sobre el problema propuesto.

- Para el uso de la herramienta de Pentaho nosotros recomendamos tener un conocimiento básico sobre bases de datos y así facilitar al usuario final la manipulación de la información.
- Para el manejo de grandes cantidades de información se recomienda que se empleen diversas herramientas visuales de software que ayudan a cubrir las necesidades de las organizaciones que manipules o manejen grandes cantidades de datos.
- Para el uso de la herramienta de WEKA nosotros recomendamos que se tenga un conocimiento básico de estadística para que el usuario final pueda interpretar los resultados, ya que en esta herramienta se emplean diversos tipos de algoritmos matemáticos que fueron enfocados para el minado de datos y pueden ser representados por medio de árboles de decisión, ayudando así a la toma de mejores opciones a diversos problemas.
- WEKA es una herramienta que recomendamos para realización de minería de datos, ya que nos ayuda a la obtención de resultados más óptimos ante un problema presentado.

Bibliografía

aaai.org. (16 de Diciembre de 2016). *aaai*. Obtenido de THE AI BEHIND WATSON — THE TECHNICAL ARTICLE: http://www.aaai.org/Magazine/Watson/watson.php

Abraha, S. H. (2006). *Fundamentos de bases de datos - Quinta Edición.* Madrid: McGraw-Hill/Interamerica de España, S.A.U.

Anónimo, P. . (14 de Mayo de 2014). *Blogspot*. Obtenido de Weka - crear un archivo .arff desde Excel: http://pathros.blogspot.com/2014/05/weka-crear-un-archivo-arff-desde-excel.html

Carnot, S. (31 de Marzo de 2014). *Hmolpedia an Encyclopedia of human thermodynamics*. Obtenido de Jeremy Camphbell: http://www.eoht.info/page/Jeremy+Campbell

CLIPS. (15 de 12 de 2016). *CLIPS*. Obtenido de CLIPS A Tool for Building Expert Systems: http://www.clipsrules.net/

CLIPS. (2 de Mayo de 2016). *CLIPS*. Obtenido de A Tool for Building Expert Systems: http://clipsrules.sourceforge.net/

Colle, R. (2002). *Explotar la información noticiosa - DATA MINING.* Madrid: Coopegraf/Visagrafic, S.L.

Concepto Definición de. (18 de Abril de 2015). Obtenido de Definición de Datos: http://conceptodefinicion.de/datos/

Davies, P. B. (1996, 2000, 2004, 2014). *Sistemas de bases de datos* . Barcelona, Bogotá, Buenos Aires, Caracas, México: Editorial Reverté.

Fundación Wikimedia, I. (8 de Marzo de 2016). *Wikipedia La enciclopedia libre*. Obtenido de Información: https://es.wikipedia.org/wiki/Informaci%C3%B3n

Gabits. (02 de Diciembre de 2009). *Algoritmos de Mineria de Datos*. Obtenido de Algoritmo "Naive Bayes": http://algoritmosmineriadatos.blogspot.com/2009/12/algoritmo-naive-bayes.html

github. (10 de Noviembre de 2016). *github*. Obtenido de Watson Developer Cloud: https://github.com/watson-developer-cloud

Google . (03 de Octubre de 2016). *Google Académico*. Obtenido de Gio Wiederhold : https://scholar.google.com/citations?user=fJiGhkoAAAAJ

Hasperué), A. A. (06 de Septiembre de 2016). *http://docplayer.es*. Obtenido de FACULTAD DE INFORMÁTICA UNIVERSIDAD NACIONAL DE LA PLATA. Tesis presentada para obtener el grado de Doctor en Ciencias Informáticas: http://docplayer.es/4602417-Facultad-de-informatica-universidad-nacional-de-la-plata-tesis-presentada-para-obtener-el-grado-de-doctor-en-ciencias-informaticas.html

Herrera, F. (06 de Septiembre de 2016). *Academia.edu*. Obtenido de Técnicas de reducción de datos en KDD. El uso de Algoritmos Evolutivos para la Selección de Instancias: http://www.academia.edu/2932699/T%C3%A9cnicas_de_reducci%C3%B3n_de_datos_en_KDD._El_uso_de_Algoritmos_Evolutivos_para_la_Selecci%C3%B3n_de_Instancias

IBM. (10 de Noviembre de 2016). *IBM*. Obtenido de Watson: http://www.ibm.com/watson/index.html

IBM. (10 de Noviembre de 2016). *IBM*. Obtenido de Watson Developer Cloud: http://www.ibm.com/watson/developercloud/

IBM. (10 de Noviembre de 2016). *IBM*. Obtenido de Watson Developer Cloud: http://www.ibm.com/watson/developercloud/starter-kits.html

IBM. (10 de Noviembre de 2016). *IBM*. Obtenido de Watson: http://www.ibm.com/watson/products.html

IBM. (10 de Noviembre de 2016). *IBM*. Obtenido de Build with Watson: http://www.ibm.com/watson/developercloud/

Iribarra, F. (06 de Septiembre de 2016). *Minería de Datos Universidad Tecnológica Metropolitana*. Obtenido de Descubrimiento del Conocimiento (KDD) : "El Proceso de minería": http://mineriadatos1.blogspot.com/2013/06/descubrimiento-del-conocimiento-kdd-el.html

Kantardzic, M. (2002). *Data Mining: Concepts, Models, Methods and Algorithms.* USA - New York: John Wiley & Sons.

KDD (Knowledge Discovery in Databases). (06 de Agosto de 2016). Obtenido de MINERIA DE DATOS Y Descubrimiento del Conocimiento: http://exa.unne.edu.ar/informatica/SO/Mineria_de_Datos_y_KDD.pdf

Klimberg, S. K. (2008). *Data Mining Methods and Applications.* Florida: Auerbach Publications - Boca Raton.

marketingdirecto.com, m. (19 de Diciembre de 2001). *md marketingdirecto.com*. Obtenido de GUÍA PARA CONSTRUIR UNA BASE DE DATOS EFICAZ: https://www.marketingdirecto.com/marketing-general/marketing/guia-para-construir-una-base-de-datos-eficaz

one, L. B. (04 de Enero de 2017). *Blog sobre Bussiness Intelligence*. Obtenido de Cómo elegir sistema de minería de datos: http://www.lantares.com/blog/como-elegir-sistema-de-mineria-de-datos

Orallo), M. M. (01 de Diciembre de 2015). *Docplayer.es*. Obtenido de Fases del KDD: Recogida de Datos. El Proceso del KDD. FASES. Fases del KDD: Recogida de Datos. Fases del KDD: Recogida de Datos: http://docplayer.es/7952056-Fases-del-kdd-recogida-de-datos-el-proceso-del-kdd-fases-fases-del-kdd-recogida-de-datos-fases-del-kdd-recogida-de-datos-proceso-detallado.html

Ortiz, A. M. (10 de 2000). *ISSN*. Obtenido de Base de Datos y Base de Conocimiento: http://elies.rediris.es/elies9/4-1.htm

Pentaho. (28 de Noviembre de 2016). *Pentaho (A Hitachi Group Company)*. Obtenido de A Comprehensive Data Integration and Business Analytics Platform: http://www.pentaho.com/

Peña, E. E. (06 de Septiembre de 2016). *http://docplayer.es*. Obtenido de T.2 Minería de Datos y Extracción de Conocimiento de Bases de Datos: http://docplayer.es/5952045-T-2-mineria-de-datos-y-extraccion-de-conocimiento-de-bases-de-datos.html

PowerData - Especialistas en Gestión de Datos. (16 de Febrero de 2014). Obtenido de El valor de la gestión de datos : http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/368784/Introducci-n-a-la-Calidad-de-Datos-Definici-n-Control-y-Beneficios

Raffo, D. (02 de Septiembre de 2009). *SearchDataCenter* . Obtenido de Técnologías para la duplicación de datos y guía práctica sobre la recuperación: http://searchdatacenter.techtarget.com/es/tutorial/Tecnologias-para-la-duplicacion-de-datos-y-guia-practica-sobre-la-recuperacion-de-desastres

Ricardo, C. M. (2004). *BASES DE DATOS.* México, D.F.: McGRAW-HILL INTERAMERICANA EDITORES, S.A. de C.V.

Sanchez, S. (07 de Mayo de 2013). *SlideShare*. Obtenido de Calidad de datos: http://es.slideshare.net/sesa78/calidad-de-datos-data-quality

Sinnexus. (07 de Diciembre de 2016). *Business Intelligence Informática estratégica*. Obtenido de ¿Qué es Business Intelligence?: http://www.sinnexus.com/business_intelligence/

Tenenbaum, A. L. (03 de Octubre de 2016). *UTM*. Obtenido de Árbol de decisión : http://www.utm.mx/~jahdezp/archivos%20estructuras/DESICION.pdf

WebMining Consultores. (10 de Enero de 2011). *WebMining Consultores*. Obtenido de KDD: Proceso de Extracción de conocimiento: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

WEKA. (28 de Noviembre de 2016). *WEKA (The University of Waikato)*. Obtenido de Weka 3: Data Mining Software in Java: http://www.cs.waikato.ac.nz/ml/weka/

Wiederhold, G. (1991). *Diseño de bases de datos.* Mexico: Programas Educativos S.A de C.V .

Wikipedia . (10 de Noviembre de 2016). *Wikipedia la Enciclopedia Libre*. Obtenido de IBM Watson: http://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/

Wikipedia . (10 de Jinio de 2016). *Wikipedia La enciclopedia libre*. Obtenido de IBM POWER: https://es.wikipedia.org/wiki/IBM_POWER

Wikipedia. (16 de Agosto de 2016). *Wikipedia (La enciclopedia libre)*. Obtenido de Análisis predictivo: https://es.wikipedia.org/wiki/An%C3%A1lisis_predictivo

Wikipedia. (05 de Septiembre de 2016). *Wikipedia (La enciclopedia libre)*. Obtenido de Aprendiaje automático: https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

Wikipedia. (05 de Octubre de 2016). *Wikipedia (La enciclopedia libre)*. Obtenido de Data dredging: https://en.wikipedia.org/wiki/Data_dredging

Wikipedia. (25 de Junio de 2016). *Wikipedia (La enciclopedia libre)*. Obtenido de Pentaho: https://es.wikipedia.org/wiki/Pentaho

Wikipedia. (17 de Mayo de 2016). *Wikipedia (La enciclopedia libre)*. Obtenido de Weka (aprendizaje automático): https://es.wikipedia.org/wiki/Weka_(aprendizaje_autom%C3%A1tico)

Wikipedia. (22 de Septiembre de 2016). *Wikipedia (La Enciclopedia libre)*. Obtenido de Conjunto de datos: https://es.wikipedia.org/wiki/Conjunto_de_datos

Wikipedia. (06 de Agosto de 2016). *Wikipedia (The free Encyclopedia)*. Obtenido de Gregory Piatetsky - Shapiro: https://en.wikipedia.org/wiki/Gregory_Piatetsky-Shapiro

Wikipedia. (18 de Noviembre de 2016). *Wikipedia la enciclopedia libre*. Obtenido de Extract, transform and load: https://es.wikipedia.org/wiki/Extract,_transform_and_load

Wikipedia. (20 de Junio de 2016). *Wikipedia La enciclopedia libre*. Obtenido de IBM POWER: https://es.wikipedia.org/wiki/IBM_POWER

Wikipedia. (17 de Octubre de 2016). *Wikipedia La enciclopedia libre*. Obtenido de Watson (inteligencia artificial): https://es.wikipedia.org/wiki/Watson_(inteligencia_artificial)

Wikipedia. (4 de Diciembre de 2016). *Wikipedia La enciclopedia libre*. Obtenido de Inteligencia empresarial: https://es.wikipedia.org/wiki/Inteligencia_empresarial

Wikipedia. (2016 de Octubre de 2016). *Wikipedia The Free Encyclopedia*. Obtenido de Watson (computer): https://en.wikipedia.org/wiki/Watson_(computer)

Wikipedia. (12 de Junio de 2016). *Wikipedia The Free Encyclopedia*. Obtenido de UIMA: https://en.wikipedia.org/wiki/Apache_UIMA

Wikipedia. (05 de Noviembre de 2016). *Wikipedia The Free Encyclopedia*. Obtenido de SUSE Linux Enterprise Server: https://en.wikipedia.org/wiki/SUSE_Linux_Enterprise_Server

Wikipedia. (10 de Octubre de 2016). *Wikipedia The Free Encyclopedia*. Obtenido de CLIPS: https://en.wikipedia.org/wiki/CLIPS

Wikipedia. (10 de Noviembre de 2016). *Wikipedia The Free Encyclopedia* . Obtenido de POWER7: https://en.wikipedia.org/wiki/POWER7

ANEXOS

1. WATSON

Watson es un sistema de tipo informático que fue desarrollado por IBM, este tiene como objetivo el contestar cualquier tipo de pregunta en lenguaje natural. Su IA es tan avanzada en el campo cognitivo por lo que es capaz de pensar e interpretar datos como un ser humano.

Watson tiene la capacidad del entendimiento y el aprendizaje del lenguaje natural que puede estar disponible y publicado en cualquier sitio web a nivel mundial, por lo que es capaz de interpretar por ejemplo tweets, artículos, informes, mensajes, publicaciones, etc.

1.1. ¿Qué es Watson?

Es una tecnología cognitiva que utiliza inteligencia artificial que es capaz de interactuar, razonar, aprender y entender cualquier tipo de información o datos, en la que se puede ver incluido texto, videos, imágenes, etc.; esto lo realiza por medio del aprendizaje de máquina.

A su vez Watson puede interactuar de forma personal con algún sujeto proporcionándole información o puede realizar recomendaciones de algún tema de interés de dicho individuo y así participar en un diálogo con el sujeto de prueba.

1.2. Descripción

Watson es un sistema que utiliza el aprendizaje de máquina el cual le sirve para aprender, razonar, responder e interactuar con grandes cantidades de información, al igual que las personas Watson por medio del lenguaje natural, puede educarse por medio de interacciones y así con cada una de las nuevas experiencias que este obtenga, puede volverse aún más inteligente en una velocidad increíblemente rápida.

Con el conocimiento obtenido de cualquier campo laboral Watson puede enseñar en dicha área como por ejemplo la medicina, economía, ingeniería, etc.; por medio de patrones realiza una búsqueda en grandes volúmenes de datos y así esta herramienta trata de proporcionándoles rápidamente mejores soluciones a diversos problemas en estas áreas. Las soluciones que genera Watson se muestra de una forma que sea compresiva para el usuario pueda entender y compartirla fácilmente con otras personas.

La información es almacena en la nube del software de IBM, la cual es accesible y cualquier momento por los usuarios de Watson.

1.1.1. Hardware

Es la integración de procesadores en paralelo de POWER7[30] con tecnología DeepQA[31]. Tiene un soporte de hardware para *Jeopardy![32]*, percibía con 90 servidores IBM POWER 75, cual cada uno de los servidores utilizaba un procesador de 3.5 GHz con un total de 8 núcleos.

Cada uno de los núcleos es capaz de soportar un hardware de 4 hilos de ejecución (threads).

En su totalidad este sistema tiene 2880 núcleos de procesamiento POWER7, y dispone de una capacidad de RAM de 16 Terabytes.

1.1.2. Software

[30] POWER7: "es un procesador de ocho núcleos capaz de correr hasta cuatro hilos en cada uno, transformado virtualmente cada procesador en un chip de 32 núcleos y así dándole una clara ventaja sobre cualquiera de los productos de Intel o AMD especializados para servidores." (Wikipedia, 2016)

[31] DeepQA: es el sistema de información y tecnología que se empleó para la elaboración de Watson, el que está encargado de responder de forma inmediata las diversas preguntas planteadas por un usuario el cual recepta la respuesta en lenguaje natural.

[32] Jeopardy!: es un concurso de Estados Unidos que tiene como finalidad el nivel de conocimiento por medio de preguntas.

El software que utiliza es Apache UIMA[33]. Este sistema fue codificado o realizado en varios lenguajes y herramientas de programación como: C++, Prolog y Java.

El sistema operativo en el que es ejecutado en SUSE Linux Enterprise Server 11, funciona conjuntamente con Apache Hadoop.

1.1.3. Watson API y SDKs

Las SDKs que están disponible para su uso conjuntamente con Watson API, pueden ser encontrada en la nube en los siguientes enlaces:

- (http://www.ibm.com/watson/developercloud/)
- (https://github.com/watson-developer-cloud)

En estos enlaces se encuentra disponible: android-sdk, python-sdk, java-sdk, ios-sdk, uniny-sdk y node-sdk.

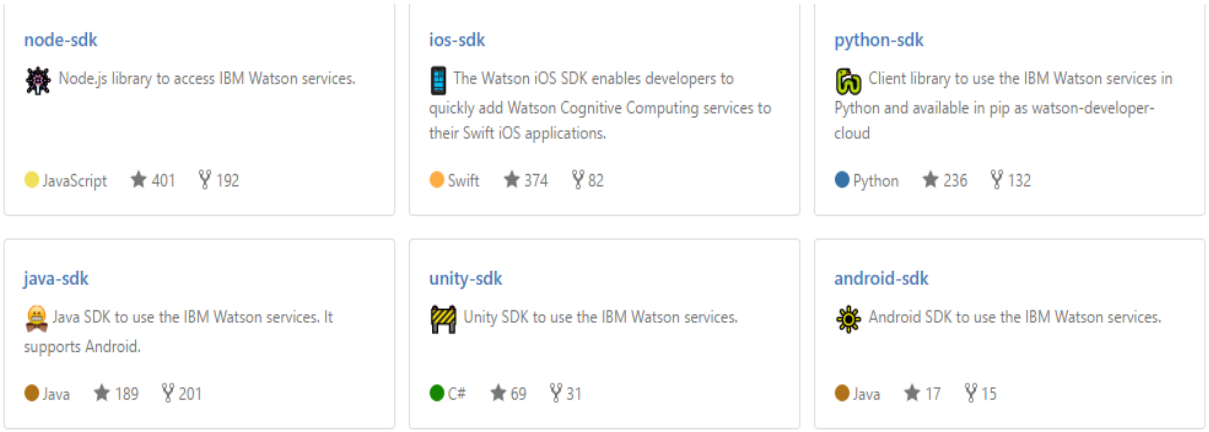


Figura Anexo 1.1. APIS para el uso de Wtason, Elaborado por: IBM (Watson)

1.1.4. Arquitectura de Watson

[33] Apache UIMA (Unstructured Information Management Architecture): es una arquitectura para la administración de información no estructura el cual puede ser empleado en la minería de texto (Text Mining) para la extracción de la información.

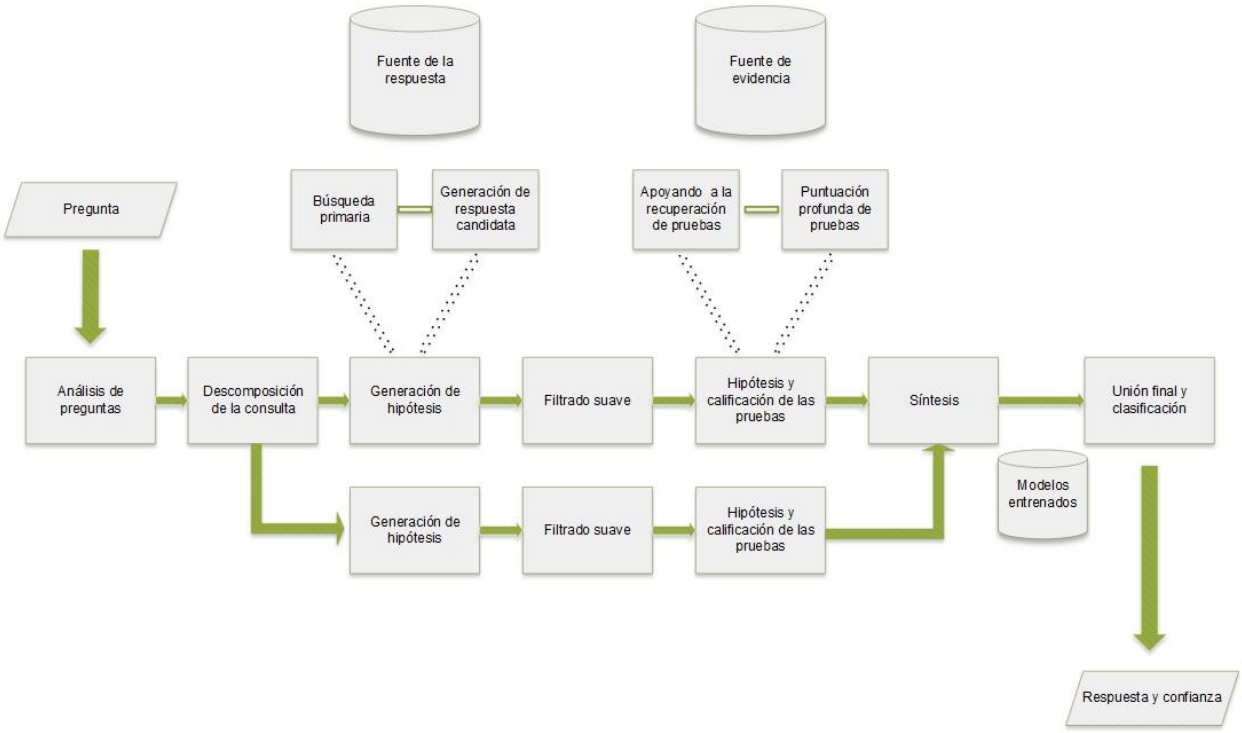


Figura Anexo 1.2. Arquitectura de Watson, Elaborado por: Oscar Córdova y Carlos Rosales

Para el entendimiento de la arquitectura de Watson que se presenta en la Figura Anexo 1.2., debemos saber que Watson es un sistema que posee una rápida integración y evaluación de grandes cantidades de información, por lo que permite generar resultados óptimos a cualquier problema planteado, en donde las respuestas deben ser de calidad, con alta precisión y pueda abordar cualquier situación.

En donde Watson se lo puede representar por medio de una arquitectura basada en datos probabilísticos que se pueden emplear de forma paralela, que por medio del lenguaje natural este pueda identificar las preguntas a cualquier tipo problema, realizando un análisis de las preguntas de dicho problema, buscándolo en varias fuentes para así descomponerlo y generar varias hipótesis, en la que las hipótesis son probadas y puntuadas para así generar la clasificación de las hipótesis. Las diversas combinaciones que se generan, aportan nuevas hipótesis o posibles respuestas que pueden formar nuevos enfoques o síntesis en los puntos más fuertes los cuales son clasificados, para la selección de una mejor solución permitiéndole contribuir a la mejora, generando calidad y confianza a los resultados en el menor tiempo de respuesta.

2. CLIPS

CLIPS ("C" Language Integrated Production System). Es una herramienta que fue creado por la NASA/Johnson Space Center desde 1985 hasta 1996, este sistema es considerado como un lenguaje de programación que es similar a "C", el cual tiene como finalidad la creación de sistemas expertos[34] que sean capaces de seguir reglas por medio de soluciones algorítmicas. Es un software multiplataforma por lo que se encuentra en constante evolución el cual puede ser empleado en cualquier sistema operativo desde el año de 1996.

2.1. ¿Qué es CLIPS?

Es un software que incorpora un completo lenguaje orientado a objetos que se emplea para la creación de varios sistemas expertos en donde se toma en cuenta la experiencia humana y hechos como factor fundamental para generar la base del conocimiento y por medio de un conjunto de reglas busca o genera una solución.

CLIPS es una herramienta de sistema experto ya que ayuda a modelar la experiencia humana tomando la información y generando mayor conocimiento facilitando así el desarrollo de sistemas expertos o software.

Un software que sea escrito en CLIPS posee un conjunto de objetos, hechos y reglas los cuales son considerados o tomados en cuenta por parte del motor de inferencia, el cual decide qué hacer y cuando deben ejecutarse las reglas que fueron basadas en dichos hechos y objetos.

2.2. Características de CLIPS

Entre las principales características para la implementación del conocimiento y el de CLIPS es necesario el tener en conocimiento las siguientes características:

[34] Sistema Experto: se basa en la experiencia y conocimiento humano el cuál sirve de referencia para la creación de diversos programas.

- Considerar la experiencia para la generación de reglas las cuales serán tomadas en cuentan al momento de representar el conocimiento heurístico.
- Tomar en cuentan las diversas funciones que van a ser usadas en para la generación del conocimiento procedimental.
- Conocer sobre la programación orienta a objetos que será usada para la generación del conocimiento procedimental.
- La generación de datos en base a los hechos e instancias.
- Tiene un conjunto de reglas que son de suma importación para la creación de la base de conocimiento.
- La manipulación y monitoreo de las reglas por medio del motor de inferencia.

2.3. ¿Con qué se integra CLIPS?

CLIPS puede ser integrado con cualquier lenguaje de programación de alto nivel como es el caso de Java, FORTRAN, Ada y C por lo que es necesario tener conocimiento básico para sacar provecho a los objetos que son empleados en el desarrollo de software.

2.4. ¿En qué se usa CLIPS?

Entre uno de los usos de este lenguaje de programación fue para la instrucción inteligente para el transbordador espacial, el cual tenía como finalidad realizar entrenamiento y prueba de los controles de vuelo, en la que los comando y las reglas de producción fueron escritas en “C”.

2.5. Versiones de CLIPS

Entre las versiones de CLIP tenemos:

- CLIPS 5.1
- CLIPS 6.24
- CLIPS 6.3

- CLIPS .NET
- CLIPS JNI
- CLIPS iOS
- CLIPS CGI